8-7-2020

Neural Correlates of Native-language Speech Perception and Non-native Speech Sound Learning

Pamela Fuhrmeister
University of Connecticut - Storrs, pamela.fuhrmeister@uconn.edu

Neural Correlates of Native-language Speech Perception and Non-native Speech Sound Learning

Pamela Fuhrmeister, Ph.D.

University of Connecticut, 2020

Many studies of non-native speech sound learning report a great deal of individual variability; some learners master the sounds of a second language with ease, while others struggle to perceive and produce sounds, even after years of learning the language. Although some contributions of phonological, auditory, or cognitive skills have been found to predict non-native speech sound learning ability as measured by laboratory tasks, the field lacks a comprehensive understanding of where these differences originate from. Recent findings, however, suggest that individual differences in sleep duration may predict learning after a period of offline consolidation, though these findings are mixed. Another issue is that the large amount of individual variability seen in studies of non-native learning makes it difficult to obtain precise estimates of effect sizes. Therefore, the first aim of this dissertation was to replicate and extend recent behavioral and neuroimaging findings in non-native speech sound learning with a larger sample size than is typical. The second goal was to test a new question, namely, that how consistently and categorically listeners perceive native-language sounds will predict success on non-native speech sound learning tasks. Finally, we sought to establish whether measures of brain structure can predict how categorically listeners perceive sounds in the native language and how consistently they respond to those sounds.

Pamela Fuhrmeister—University of Connecticut, 2020

We did not replicate recent findings showing behavioral improvement after sleep on non-native speech sound learning tasks, nor did we replicate the finding that sleep duration predicts overnight improvement. However, gyrification of the bilateral transverse temporal gyri and hippocampal volume predicted an individual's overnight improvement, suggesting a role for memory consolidation, even though we did not see overnight improvement at the group level. We additionally did not find that individual differences in categorical perception predicted non-native speech sound learning, which presents a challenge for some predominant theories of non-native speech sound learning, which future research will have to address. Overall, learners with reduced surface area and volume in frontal regions showed more graded and consistent perception of native-language speech sounds, supporting the notion that these regions underlie categorical perception.

Neural Correlates of Native-language Speech Perception and Non-native Speech Sound Learning

Pamela Fuhrmeister

B.A., Texas Tech University, 2009

M.A., Texas Tech University, 2011

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2020

APPROVAL PAGE

Doctor of Philosophy Dissertation

Neural Correlates of Native-language Speech Perception and Non-native Speech Sound Learning

Presented by Pamela Fuhrmeister, B.A., M.A.

Major Advisor _____
Emily Myers, Ph.D.

Associate Advisor _____
Betsy McCoach, Ph.D.

Associate Advisor _____
Erika Skoe, Ph.D.

Associate Advisor _____
Rachel Theodore, Ph.D.

University of Connecticut

2020

ACKNOWLEDGMENTS

I am truly overwhelmed and humbled when I think about all the teachers, mentors, and many other supportive people in my life who have helped me along the way to complete my PhD. First and foremost, I will always be grateful to my advisor, Dr. Emily Myers for taking a chance on someone who came to the table with little more than a few burning questions about speech perception. Emily helped me take my ideas and turn them into interesting, theoretically motivated questions, and she consistently reminded me not to lose sight of the big-picture implications of my work. Her trust in me as a scientist, even from the very beginning, helped me gain confidence and independence. Though Emily has high standards for her lab, I never once felt discouraged. Her direct, but gentle and often humorous delivery of feedback would leave me feeling inspired to learn more and make my work better. Perhaps most importantly, Emily has created a rigorous, but fun and collegial lab environment, where we take our work very seriously, but we don't take ourselves too seriously. It was a joy to come to work every day for the last five years, and I hope to create a similar lab environment in the future.

My graduate education would not have been complete without the strong community of language researchers at the University of Connecticut. I have greatly benefitted from the mentorship and guidance of my committee members and readers, Dr. Rachel Theodore, Dr. Erika Skoe, Dr. Betsy McCoach, and Dr. Adrian Garcia-Sierra. Each of them has been so generous with their time and advice they have given me about science or careers. I know that this project is better because of their input, and I am a better researcher because of what I have learned from them.

This project would not have been possible without the help of several people. I am grateful to Elisa Medeiros for her impressive expertise, for helping nervous MRI participants feel comfortable and be able to complete the study, and mostly for making the many hours we spent together scanning participants for this study a lot of fun. I am also grateful to Dr. Peter Molfese for his help with the gyrification analyses for this project, for teaching me about MRI, and for reassuring me that life gets better after your dissertation is done. I would also not have been able to complete this project without Hannah Mechtenberg's help running participants when I was gone, organizing everything in the lab, and training a great group of undergraduate research assistants who also helped with this project.

I want to thank the past and present members of the Language and Brain Lab, especially Sahil Luthra, Dave Saltzman, Dr. Kathrin Rothermich, Dr. Chris Heffner, and Hannah Mechtenberg. My work is better because of their feedback on so many presentations and manuscript drafts. I am grateful they shared their knowledge and expertise with me and were willing to spend sometimes hours at a time helping me work through code errors or just think through ideas. I have learned so much from all of them, and their friendship and collegiality made it fun to come to the lab everyday. I also want to thank Dr. Julia Drouin for her friendship and support, and for encouraging me to "just set a date!" for all my PhD milestones. I am glad we did our milestones together, and I will miss our Starbucks writing dates.

I am grateful to my family for the opportunities they provided me that have enabled me to pursue this path. My dad showed me what hard work looks like and has made countless sacrifices to help me get an education. Even when money

was tight, he never hesitated to do whatever he could to help me with something education-related. My brother, Tommy, has probably been my biggest cheerleader ever since I can remember. It means so much to me that he expresses genuine interest and enthusiasm for my work, even the boring details, and shares my joy for all my PhD and other life milestones.

Finally, I want to thank my husband, Garrett. Through all the ups and downs of doing a PhD (and everything else life throws at us), his unconditional love, support, and encouragement has been constant. Knowing he believed in me gave me the courage to even apply to PhD programs and to stick with it when it was difficult. Most of all, Garrett reminds me that work is just one important aspect of our lives and that spending time with the people we love is really what matters.

Contents

Chapter 1

Introduction

Learning a second language in adulthood can be a challenging process, and individuals vary substantially in their ultimate success (e.g. Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; Flege, Munro, & MacKay, 1995; Flege, Yeni-Komshian, & Liu, 1999). Much of the variability seen in ultimate outcomes in second language learning can be attributed to factors that are fairly easily observed, such as age of acquisition, time spent in a country or relative time spent using the second language, or motivation to improve proficiency (see Flege et al., 1995, 1999; Piske, MacKay, & Flege, 2001). After some time spent learning a second language, many adult learners achieve at least moderate success in some domains of language acquisition, such as syntax or word learning (e.g. Flege et al., 1999; Granena & Long, 2013). However, a common but perplexing finding persists in this literature: Adult learners exhibit a wide range of variability in their ability to learn to perceive and produce the speech sounds of a second language (e.g. Bradlow et al., 1999; Lim & Holt, 2011; Myers &

Swan, 2012; Yi, Maddox, Mumford, & Chandrasekaran, 2016). In fact, substantial individual variability can be found in almost all published studies on non-native speech sound learning. Although many studies have sought to elucidate the sources of individual variability in non-native speech sound learning, as a field, we still have a very incomplete picture of why some individuals excel at this skill and why some struggle.

A less-obvious question is whether individuals demonstrate variability in measures of native-language speech perception or ability. As discussed in more detail below, listeners vary in how categorically they perceive speech sounds (Kapnoula, Winn, Kong, Edwards, & McMurray, 2017; Kong & Edwards, 2016) and other measures of phonological skills and phonological working memory (Earle & Arthur, 2017; Fuhrmeister, Schlemmer, & Myers, accepted). Some of these skills have been found to predict non-native speech sound learning in children and adults (Earle & Arthur, 2017; Fuhrmeister et al., accepted; MacKay, Meador, & Flege, 2001; Perrachione, Lee, Ha, & Wong, 2011). Much work on differences in native-language speech perception has focused on understanding how these skills differ in groups with language or reading disorders (e.g. Serniclaes, Van Heghe, Mousty, Carré, & Sprenger-Charolles, 2004; Werker & Tees, 1987), but emerging evidence suggests that individual variability in native-language speech perception is present even in a typical population (e.g. Kapnoula et al., 2017; Kong & Edwards, 2016).

An open question is whether individual differences in non-native speech sound learning can be explained by individual differences in native-language speech measures. In the first part of this chapter, I will review studies examining possible sources of

individual variability in native and non-native speech processing, and I will argue that individual variability in categorical perception of native-language speech sounds should predict individual variability in non-native speech sound learning. To better understand sources of individual variability in native and non-native speech processing, the second part of the chapter looks to structural and morphological variability in brain structure as a source of individual differences in behavior. Finally, the chapter concludes with an appeal to the field to obtainmore accurate estimates of effect sizes (a goal of this dissertation) because the individual variability so commonly observed in non-native speech sound learning research makes it difficult to draw reliable and replicable conclusions from the existing published data. Elucidating these sources and consequences of variability in various measures of speech perception may inform our understanding of typical and atypical language processing, as well as language learning.

1.1 Categorical perception

Perhaps the most well-known finding in the field of speech perception is that listeners perceive speech sounds categorically (e.g. Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman, Harris, Hoffman, & Griffith, 1957). Evidence for this phenomenon comes from studies in which listeners categorize and discriminate speech sounds taken from a synthetically modified continuum from, for example, /dɑ/ to /tɑ/, in which voice onset time is increased in equal steps to form the continuum. Categorization data with such a continuum typically reveal a sharp increase in the

proportion of /tɑ/ responses near the category boundary, rather than a gradual increase as VOT increases. Complementing categorization data is discrimination data that often show poor discrimination of tokens within a category but relatively good discrimination for tokens that span a category boundary.

1.1.1 Graded vs. categorical perception

Many early studies assessed categorical perception using two-alternative forced choice (2AFC) tasks (e.g. Liberman et al., 1967, 1957). Although findings of categorical perception of speech using this method are robust, we have known for some time from studies using more sensitive measures than a 2AFC task that listeners do not discard the within-category acoustic-phonetic variation and in fact maintain sensitivity to subtle within-category differences. For example, studies utilizing reaction time data (Pisoni & Tash, 1974), goodness judgments (Drouin, Theodore, & Myers, 2016; Miller, 1997), eye tracking (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; McMurray, Danelz, Rigler, & Seedorff, 2018; McMurray, Tanenhaus, & Aslin, 2002), and visual analog scaling tasks (Kapnoula et al., 2017; Kong & Edwards, 2016) have found that listeners can indeed distinguish subtle within-category differences in speech stimuli and this might confer certain advantages when understanding spoken language. For example, the ability to detect subtle acoustic detail in the speech signal can help a listener recognize words by anticipating coarticulation (Gow, 2001). A common example of this is the difference in the /s/ sound in the words “see" and “sue." The acoustic properties of these two /s/ sounds differ depending on the following vowel, and sensitivity to these acoustic differences may help a listener predict upcoming speech

sounds and in turn, recognize words more quickly. Sensitivity to subtle differences in the speech signal can also help a listener track a particular distribution of a speech sound (Clayards et al., 2008).

Many studies that have found more graded speech representations among adults have not necessarily looked at individual variability in how graded or categorically an individual perceives speech. However, a few recent studies using visual analog scaling tasks (a task in which participants are asked to move a slider along a continuum to indicate where they think a token lies between, in this case, two speech sounds) suggests that even typically developing adults vary in how categorically or graded they perceive speech sounds (Kapnoula et al., 2017; Kong & Edwards, 2016). Both of these studies used a visual analog scaling task to test how categorically individuals perceive speech sounds. The earlier study by Kong and Edwards (2016) provided some of the first evidence that a visual analog scaling task could measure individual differences category gradiency, and the study by Kapnoula et al. (2017) validated this technique with a substantially larger sample size of over 100 participants. In addition, the study by Kapnoula et al. (2017) used a novel statistical approach to measure integration of a secondary acoustic cue. Their results showed that individuals vary substantially in how categorically or graded they perceive consonants along a voice onset time continuum, and they showed that participants who showed more graded perception of the sounds were more successful at integrating a secondary acoustic cue to distinguish voiced and voiceless stop consonants (F0). It is not clear whether graded perception causes better integration of secondary acoustic cues; however, this study provides evidence that they are related, and secondary cue integration is

important for perceiving many distinctions among speech sounds. Therefore, this study adds to the mounting evidence that graded perception is beneficial for the listener.

1.1.2 Limitations of tasks used to measure categorical perception

Although 2AFC tasks were common in early studies testing categorical perception (e.g. Liberman et al., 1967, 1957), more recent studies provide evidence that they may not be an optimal measure of gradedness of a category representation, and shallower categorization slopes may be more indicative of noisy representations. For example, if a listener perceives speech sounds in a graded manner, they can perceive subtle, within-category differences in sounds. This means they will likely have a very precise boundary between two categories, so when presented with a 2AFC task, these graded listeners will likely have very accurate and consistent responses according to their category boundary, which will result in steep categorization slopes. Although steeper categorization slopes have often been taken as evidence of categorical perception, more graded representations may actually lead to more categorical-like response patterns in 2AFC tasks due to the precision these listeners have for their category boundaries (e.g. McMurray et al., 2002). On the other hand, if a listener cannot perceive subtle within-category differences in speech sounds, their category boundaries may be less precise, and when tokens fall close to the boundary, these listeners may be less accurate or less consistent in responding to those tokens on a 2AFC task. This would result in a flatter categorization function that has historically been taken as evidence of

graded perception, such as in children in the study by Burnham, Earnshaw, and Clark (1991). Instead of interpreting flatter categorization functions as evidence of graded perception, we may need to start thinking of them as evidence of noisy representations (due to less precise category boundaries) or simply less reliability of responses (i.e., young children may be less reliable responders). Experimental evidence supporting this notion can be found in recent work comparing performance on 2AFC tasks with other, more sensitive measures of perception. For instance, McMurray et al. (2018) tested children on a phoneme categorization task and also obtained eye-tracking measurements of lexical competition with tokens that varied along VOT and fricative continua and found that younger children indeed had shallower categorization slopes (which have traditionally been taken as evidence of more graded representations). At the same time, eye-tracking data revealed the opposite pattern: older children looked to the competitor item more often as tokens came closer to the category boundary, where younger children did not. These results are important for a few reasons. First, they suggest that perception of speech sounds becomes more graded throughout development, and this is further support for the notion that graded perception may be indicative of a mature representation of speech sounds. Second, they indicate that measuring how categorically or graded an individual perceives sounds is complicated and depends on the task involved. In addition, Kapnoula et al. (2017) found no relationship between an individual's categorization slope from a visual analog scaling task (which allows for more graded responses) and the slope from a 2AFC task, supporting the idea that these two tasks measure different aspects of speech perception. On the whole, it seems that most adults perceive speech

sounds less categorically than originally thought and that graded perception may reflect a mature category representation that supports spoken language processing. Furthermore, 2AFC tasks that have been used to demonstrate categorical perception seem to limit listeners' ability to show how graded their perception of speech sounds is. For example, a listener who has a clear boundary between two speech categories may be able to distinguish within-category speech tokens; however, this listener would also be able to recognize which category a near-boundary token would belong to (i.e., that a /dɑ/ with a voice onset time of 25ms is different than one with a voice onset time of 10ms but still a better exemplar of /dɑ/ than /tɑ/). This suggests that future studies investigating individual or group differences in how categorically speech sounds are perceived should use other, more sensitive measures than a 2AFC task.

If graded perception is advantageous, it is important to understand the consequences when listeners are not able to perceive these subtle acoustic-phonetic differences in speech sounds. Some of these consequences may be subtle because there are many redundant speech cues and many contextual cues available to aid comprehension. However, as explained in the next section, a lack of sensitivity to within-category speech differences may hinder learning of new speech categories, as one does when learning speech sounds in a second language.

1.2 Relationships between native and non-native speech perception

1.2.1 Predictions from theories of non-native speech sound learning

Prevalent theories of non-native speech sound learning make clear predictions about which speech sounds will be most difficult to learn for an individual with a certain native-language background (e.g., the perceptual assimilation model, Best, McRoberts, & Goodell, 2001; Best & Tyler, 2007, the native-language magnet model, Kuhl, 1994; Kuhl et al., 2008, and the attention to dimension model Francis & Nusbaum, 2002). These theories predict that perceptual similarity of native and non-native speech sounds determines how difficult a given speech sound will be for a learner. For example, native speakers of English struggle to discriminate voiced dental and retroflex stop consonants found in Hindi due to their perceptual similarity to the English voiced alveolar stop consonant (e.g. Werker & Tees, 1984), and because some allophonic variants of these sounds occur in English (e.g., in “width" or “drip," Polka, 1991). However, native English speakers often have little trouble discriminating Zulu click sounds because there are no similar speech sounds found in English (Best, McRoberts, & Sithole, 1988). Each of these theories differ slightly in the details. For instance, the native-language magnet model posits that during development, children develop speech category prototypes based on the acoustic input they receive. These prototypes then warp perceptual space, and as a result, these category prototypes attract any

perceptually similar speech sounds like a magnet, which makes it difficult to perceive certain non-native sounds as their own distinct categories (Kuhl, 1994; Kuhl et al., 2008). The perceptual assimilation model assumes that articulatory gestures underlie perception, and that perceptually similar non-native speech sounds will be assimilated to native-language speech categories that have a similar articulatory gesture (Best et al., 2001; Best & Tyler, 2007). Attention to dimension models focus more on the cognitive processes involved with learning non-native speech sounds. Specifically, this model posits that native speakers of a language have learned to direct their attention to certain dimensions of the acoustic signal, depending on which acoustic cues are relevant and informative in the native language. Learning a new sound simply requires the learner to direct attention to a new set of acoustic cues (Francis & Nusbaum, 2002). Despite their differences, these theories have one overarching principle in common: Perceptual proximity to native-language speech sounds is the main source of difficulty in non-native speech sound learning.

Although these theories predict fairly accurately which specific speech sounds will be most difficult to learn for a speaker of a given language, a challenge for them is that most studies of non-native speech sound learning report a great deal of individual variability among learners of the same first-language background. None of these theories make explicit predictions about an individual learner; however, we can infer from their predictions why some individuals may experience more or less difficulty learning challenging non-native speech sounds. If learners are unable to discriminate non-native speech sounds because they assimilate them to perceptually similar native-language speech categories, it is logical to assume that learners who

are less likely to assimilate new sounds onto existing categories will have an easier time learning non-native speech sounds. Listeners who demonstrate more graded perception of native-language speech sounds (in other words, they are sensitive to subtle, within-category differences) may be less likely to assimilate perceptually similar speech sounds to existing categories, as subtle acoustic differences in the speech signal may be more salient to these individuals. Thus, perception of native-language speech sounds that is less categorical (and therefore more graded) may be advantageous for learning challenging non-native speech sounds. A finding in which more graded perception of native-language speech sounds positively predicted non-native speech sound learning would lend support to these theories (i.e., that perceptually assimilating non-native speech sounds to native-language categories is what makes non-native speech sound learning difficult).

A challenge for this idea can be found in studies examining children's perception of native-language phonological structure and their non-native speech sound learning abilities. One study found that children did not assimilate non-native speech sounds to native-language categories as much as adults did (Baker, Trofimovich, Flege, Mack, & Halter, 2008), and it is widely known that outside the laboratory, children often master the speech sounds of a second language more successfully than adult learners (e.g., Flege et al., 1995, 1999; Granena & Long, 2013; Piske et al., 2001). Children's perception of speech sounds has been argued to be less categorical than adults' perception as indicated by shallower categorization slopes from a 2AFC task (Burnham et al., 1991; Hazan & Barrett, 2000), which would be consistent with the idea that graded perception results in better learning of non-native speech sounds.

However, as discussed above, McMurray et al. (2018) found the opposite pattern using eye tracking (a more sensitive measure than a 2AFC) and found that children's perception becomes more graded throughout adolescence. Taken together, these findings present a challenge for theories of non-native speech sound learning because it is not consistent with the idea that graded representations of native-language speech sounds predicts better non-native speech sound learning. However, it is possible that more graded representations confer an initial advantage in learning non-native speech sounds but that other learning or memory processes contribute to long-term learning (Fuhrmeister et al., accepted, see Fuhrmeister, 2019, for review).

1.2.2 Similarities and differences of native and non-native speech

Processing of native and non-native speech shares many similarities. For example, in either a native or non-native language, a listener has to first process incoming acoustic information, map this information to phonemes, determine where word boundaries are, activate lexical candidates and choose the appropriate one (e.g., TRACE, McClelland & Elman, 1986), and finally, derive meaning from the words at the sentence level. Doing this requires integrating top-down and bottom-up information (Norris, McQueen, & Cutler, 2003). Both native and non-native listeners are able to take advantage of top-down information to aid comprehension in degraded listening conditions (though non-native listeners require a clearer bottom-up signal to be able to take advantage of top-down information, Bradlow & Alexander, 2007).

Despite the many similarities in learning and processing native and non-native

speech, there are clear differences, especially concerning learning. First, most people learning a spoken language begin learning their first language in infancy, while many second-language learners begin learning at a later age, and the brain may be more plastic earlier in life and may be better able to learn speech sounds at a younger age (see Werker & Hensch, 2015 for review). Additionally, an adult who is learning a new set of speech categories has the problem that the native-language categories may interfere with learning new ones (e.g., Best & Tyler, 2007). That is, an adult learner cannot approach learning of speech sounds in the same way a first-language learner can because the native-language speech categories may influence how non-native sounds are perceived (see section above on theories of non-native speech sound learning). Furthermore, it is possible that children may rely on different (perhaps more optimal) mechanisms to learn non-native speech sounds or they may have advantages in memory consolidation of non-native speech sounds (Fuhrmeister, 2019; Fuhrmeister et al., accepted). In short, there are many similarities between understanding speech in a native and non-native language, but there are crucial differences, especially in learning of those sounds, that should be taken into consideration when making assumptions about the links between native and non-native speech perception.

1.2.3 Experimental evidence for links between native and non-native speech processing

Several studies have found links between native and non-native speech processing. For example, Díaz, Baus, Escera, Costa, and Sebastián-Gallés (2008) carried out an ERP study and found that the mismatch negativity component (an electrophysiological

component that detects change in an auditory stimulus in an oddball task) was greater in good perceivers than poor perceivers when detecting a stimulus change for both native and non-native speech contrasts. Interestingly, they found no group differences with non-speech auditory stimulus changes, leading them to conclude that a speech-specific mechanism underlies perception of native and non-native speech sounds. Several other studies have used standardized measures of native-language phonological skills to predict non-native speech sound learning and have found that some of these measures predict performance on tasks of non-native speech sound learning or perception (Earle & Arthur, 2017; Fuhrmeister et al., accepted; MacKay et al., 2001; Perrachione et al., 2011). However, none of these studies have specifically looked at categorical perception as an individual difference measure of native-language speech perception. Nonetheless, these studies point to a relationship between native and non-native speech abilities that warrants further investigation.

1.3 Individual differences in brain structure

Individual differences in speech perception abilities must originate somewhere, and individual variability in brain structure is a logical place to look for potential brain-behavior relationships especially because changes in brain structure take place much more slowly than changes in functional brain activity (see Golestani, 2014, for review). Brain structure is especially interesting to look at in relation to behavior because differences in brain structure can arise from experience but some aspects of brain structure are thought to be innate. According to the radial unit hypothesis, the

cortical surface develops as a result of progenitor cells in the ventricular zone that migrate radially guided by radial glial cells, which forms the columnar organization of the cerebral cortex (Rakic, 2000). This process of cortical surface development and therefore the gyrification patterns (the folding patterns of the cerebral cortex), takes place in utero between the 6th and 20th weeks of gestation (Rakic, 2000; White, Su, Schmidt, Kao, & Sapiro, 2010). Gray and white matter, on the other hand, are more malleable and can change with experience. Therefore, individual differences in measures of gray and white matter structure may reflect experience-based plasticity, while gyrification patterns may reflect genetic differences. Additionally, the exact regions where we see variability in brain structure may give us clues as to what common mechanism underlies individual differences in (non-) native speech perception abilities. While several studies have explored relationships between brain structure and non-native speech sound learning or perception (as explained in more detail in the next section), no studies to our knowledge have sought to find anatomical correlates of individual differences in native-language speech perceptual abilities, at least in typical populations. The purpose of this section is to review the literature on individual differences in brain anatomy as it relates to speech abilities to generate predictions about how they may relate to individual differences in perception of native-language speech sounds.

1.3.1 Morphological variability of Heschl's gyrus

The size and gyrification patterns of Heschl's gyrus vary among individuals. Common morphological variations include split, duplicate, and sometimes even multiple Heschl's

gyri (Marie, Maingault, Crivello, Mazoyer, & Tzourio-Mazoyer, 2016). A large-scale study with over 200 right-handed participants found that approximately 64% of the sample had split or duplicate Heschl's gyri in either the right or left hemisphere (Marie et al., 2016). Heschl's gyrus is of interest to speech research because it contains primary auditory cortex. Indeed, several studies have found that variation in gyrification patterns in Heschl's predict non-native speech sound learning ability, musical ability, and phonetic expertise. Golestani, Molko, Dehaene, LeBihan, and Pallier (2007) found that faster learners of a challenging non-native phonetic contrast were more likely to have multiple or split Heschl's gyri in the left hemisphere. Turker, Reiterer, Seither-Preisler, and Schneider (2017) obtained similar findings: Participants who scored higher on a Hindi speech sound imitation task were more likely to have multiple or split Heschl's gyri, but in the right hemisphere. Interestingly, split or duplicate Heschl's gyri have been linked to both phonological deficits (Leonard et al., 2001) and phonetic expertise (Golestani, Price, & Scott, 2011). Leonard et al. (2001) found that individuals with dyslexia who exhibited a phonological deficit were more likely to have multiple or split Heschl's gyri compared to a group of typical readers. However, Golestani et al. (2011) found that a group of expert phoneticians had more occurrences of multiple or split Heschl's gyri compared to a group of controls with a comparable education background. Findings from these two studies are difficult to reconcile, but neither study had a large sample size. For instance, the study by Leonard et al. (2001) had only 11 participants in the group with phonological dyslexia, and the study by Golestani et al. (2011) only had 17 in their group of expert phoneticians. As discussed more below, low-powered studies due to small sample sizes

can result in overestimates of effects or even effects that have the wrong sign (Gelman & Carlin, 2014). Therefore, it is not surprising that we see conflicting evidence in the MRI literature, where many studies suffer from small sample sizes (see Button et al., 2013, for review). In addition, the criteria for phonological dyslexia in the study by Leonard et al. (2001) was determined by a pseudoword decoding task. It is possible that this skill is different from skills that the group of phoneticians in the Golestani study had acquired from their phonetic training. Another possibility is that people with phonological dyslexia are actually adept at distinguishing subtle differences in speech sounds, and this ability to detect subtle differences in sound is related to gyrification patterns in primary auditory areas. This would be consistent with Serniclaes' theory of allophonic perception in dyslexia (Serniclaes et al., 2004), which posits that people with dyslexia often perceive allophonic variants of speech categories as separate sounds, which makes it difficult to map the sounds to a common grapheme. Overall, we lack the evidence to conclude whether multiple or split Heschl's gyri are predictive of disordered phonological processing or phonological expertise, or whether this reflects two different skills. On the whole, it seems that individuals who have more occurrences of multiple or split Heschl's gyri are better at detecting subtle differences in sounds, whether those sounds are non-native speech sounds or native-language speech sounds. Because gyrification patterns are established very early in development, this may suggest that some individuals have a predisposition for attending to subtle differences in sound. From these studies, we can predict that listeners who have higher instances of split or duplicate Heschl's gyri will perceive native-language sounds in a more graded manner because graded perception relies on

the ability to distinguish very minor differences in speech sounds.

1.3.2 Anatomical correlates of non-native speech sound learning and perception

Several studies have explored structural brain differences to explain individual variability of non-native speech sound learning, and many have found structural variability in typical language regions, especially auditory areas, that relate to this skill. Using voxel-based morphometry, Golestani, Paus, and Zatorre (2002) found greater white matter density in a region anterior to the parieto-occipital sulcus in faster learners than in slower learners of the Hindi dental/retroflex contrast. Of note, however, was that the groups did not differ in their final posttest scores, but rather on their learning rate. Golestani and colleagues (2007) found greater white matter densities in Heschl's gyrus for faster vs. slower learners of the Hindi dental/retroflex contrast. Wong et al. (2008) found greater gray matter densities in Heschl's gyrus in more successful learners of a non-native tonal contrast. In many of these studies, structural or morphological variability in primary auditory cortex predicts sensitivity to or learning of non-native (supra-)segmental contrasts. Because early auditory areas typically process fine-grained acoustic details of a stimulus, these findings suggest that individual variability in non-native speech sound learning may in part be explained by one's ability to attend to subtle acoustic details in the speech stream.

Other studies have found anatomical variation in frontal regions that relate to non-native speech perception or learning. Sebastián-Gallés et al. (2012) found differences between good and poor perceivers of native and non-native vowels in

a region encompassing the right insula and frontal operculum. Specifically, poor perceivers had more white matter density in this region, which the authors interpreted as possibly resulting from the use of compensatory strategies in speech perception (i.e. greater reliance on frontal regions as opposed to sensory areas in temporal regions). However, participants in this study were bilingual. Rodriguez, Archila-Suerte, Vaughn, Chiarello, and Hernandez (2018) found somewhat similar results, namely that cortical thickness of the left insula predicted non-native speech sound learning in bilinguals, but this pattern did not hold for monolinguals. Therefore, the extent of the relationship between the structure of the insula and non-native speech sound learning or perception is unclear.

1.3.3 Structural and functional architecture of phonological category structure

Few studies have directly examined relationships between brain structure and individual differences in native-language speech perception (specifically categorical perception of native-language speech categories); however, we can make some predictions about where these differences might emerge from the functional activation literature. Well-established findings from functional MRI studies indicate that the left superior temporal gyrus and left inferior frontal gyrus are some of the main brain regions involved in processing native-language speech (e.g., Damasio & Geschwind, 1984; Price, 2012). The brainstem encodes stimuli with high fidelity (Bidelman, Moreno, & Alain, 2013; Skoe & Kraus, 2010), but at some point in the auditory processing stream, these sounds are perceived categorically. Using a variety of methods,

evidence from several studies suggests that frontal and temporal regions underlie representations of phonetic category structure.

Many studies have found that posterior regions, such as the superior temporal gyrus, is involved in categorical perception. Some evidence suggests that categorical perception emerges in secondary auditory cortex including the posterior superior temporal gyrus (Bidelman et al., 2013; Chang et al., 2010). The study by Chang and colleagues (2010) used electrocorticography (a technique in which electrodes are placed directly onto the cortical surface in patients undergoing brain surgery) and found that parts of the superior temporal gyrus respond invariantly to specific acoustic-phonetic features. An fMRI study by Myers (2007) found that the superior temporal gyrus responds to speech category structure in a graded manner. Specifically, greater activation was found in bilateral superior temporal gyri when tokens from a stop continuum that were poor members of the category (either exaggerated stimuli that were not competitive with another category or near-boundary tokens that were competitive with another category) were heard. This suggests that this region is not only sensitive to the category boundaries of speech sounds, but it is also sensitive to how prototypical a given exemplar is of the category it is perceived as. Functional activation in left temporal areas has also been found to predict individual differences in categorization of phonemic and non-phonemic stimuli. An fMRI study by Desai, Liebenthal, Waldron, and Binder (2008) suggests that a region encompassing the left posterior superior temporal gyrus and left posterior superior temporal sulcus is more active in response to sine wave speech when participants perceive the tokens as speech as compared to before participants are aware of the phonemic properties of the

stimuli. In addition, activation in this region predicted how categorically participants perceived the speech and non-speech continua. Therefore, we may see that individual differences in how graded or categorically sounds are perceived may be correlated with differences in brain structure or morphology in left superior temporal areas.

In addition to temporal regions, several studies of native-language speech perception and non-native speech sound learning have suggested a role for frontal regions in categorical perception as indicated by changes in activation for members of different phonetic categories but no change in activation for acoustically distinct members of the same category. fMRI studies using univariate and multivariate approaches have found that the left inferior frontal gyrus and left middle frontal gyrus have been shown to respond more categorically, or invariantly, to speech categories (Lee, Turkeltaub, Granger, & Raizada, 2012; Myers, 2007; Myers, Blumstein, Walsh, & Eliassen, 2009; Myers & Mesite, 2014; Myers & Swan, 2012). As just discussed, Myers (2007) found that bilateral inferior frontal gyri show greater activation for stimuli near a category boundary, suggesting these regions may help resolve competition among competing alternatives. Myers et al. (2009) found that the left inferior frontal sulcus responded invariantly to speech sounds. Using a multivariate analysis approach, Lee et al. (2012) found a similar pattern of results showing that Broca's area of the left inferior frontal gyrus showed patterns of activation consistent with categorical representations in two different data sets. Myers and Mesite (2014) and Myers and Swan (2012) show evidence that the middle frontal gyri show categorical-like responses in perceptual learning tasks or newly learned phonetic categories. Because brain function and brain structure are often related, it is likely that individual differences in brain structure

that relate to behavioral differences in categoricity will be found in these areas or nearby, functionally related areas.

One study (also discussed in previous sections) of individual differences in brain structure and morphology hints at some relationships between the brain and individual differences in phonetic category structure. Golestani et al. (2011) looked for anatomical differences between a group of expert phoneticians and a group of non-expert controls and found that expert phoneticians were more likely to have multiple or split Heschl's gyri compared to the controls. Additionally, gray matter volume of the pars opercularis, a region in the inferior frontal gyrus, was predicted by years of phonetic training. Although this study did not test the participants' perception of native-language phonetic category structure, it is nonetheless interesting that both frontal and temporal regions predicted phonetic expertise. Therefore, it is possible that brain structure may differ as a function of native-language speech ability or perception of native-language speech sounds. If graded perception of speech sounds indeed represents a mature or optimal representation of speech sounds, and phonetic expertise is predicted by differences in structure and morphology of Heschl's gyrus and the inferior frontal gyrus, we expect those regions to also be related to how graded or categorically an individual perceives speech sounds in the native language.

1.4 Replicability of non-native speech sound learning findings

Replicability of findings has recently become a topic of interest in psychology and related fields. In 2015, there was a large effort to replicate 100 studies published in three different psychology journals (Open Science Collaboration, 2015), and they found significant results in only about a third of the replication attempts. One possible reason for these failures of replication is that many studies include small and noisy samples. When statistically significant results are obtained from small, noisy samples, they are necessarily inflated (errors of magnitude) or sometimes even go in the wrong direction from the true effect (errors of sign, Gelman & Carlin, 2014). In addition, p-values are rarely replicable, even when data from replication or similarly designed studies show similar patterns of results (Amrhein, Korner-Nievergelt, & Roth, 2017). Other reasons that studies may not replicate include p-hacking, HARKing (hypothesizing after results are known), and publication bias (Bishop, 2019). All of these factors combined result in an entire field basing conclusions off of low-powered studies whose effect sizes may or may not be accurate estimates. For some study designs, it is not possible to recruit an appropriate sample size. Studies of special populations, where the population of interest may be hard to recruit; longitudinal studies that have a high attrition rate; or neuroimaging studies that are expensive to run are especially susceptible to these concerns. When this is the case, it is extremely important to base conclusions off of several studies, rather than only a few positive findings that have been published, and replication efforts can help minimize this bias.

The topic of replication is extremely relevant for research in non-native speech sound learning and neuroimaging for several reasons. First, almost all published studies of non-native speech sound learning report a large amount of individual variability among the participants who were tested. Second, many of these studies have very small sample sizes (many under 30 participants per group). This means that many of the findings the field takes for granted may have been found from small, noisy samples, and therefore, the probability is high that many of these findings are errors of magnitude or sign (Gelman & Carlin, 2014). As we have alluded to in this chapter, some findings in the field are difficult to reconcile or are downright contradictory, especially in the neuroimaging literature. In fact, many neuroimaging studies are underpowered, and this makes it harder to detect a true effect if one is indeed there, but more importantly, effects that are detected are necessarily less reliable (Button et al., 2013). Therefore, one of the goals of this dissertation is to try to conceptually replicate some of the findings in the non-native speech sound learning literature with a larger than typical sample size, especially for structural MRI studies, to get closer to accurate estimates of effect sizes than have previously been reported in the literature.

1.5 Conclusion

For decades, researchers have observed that learning non-native speech sounds is subject to a wide range of individual variability. Exactly where this variability stems from is still an open question, and both theoretical predictions and experimental

evidence suggest that individual differences in the perception of native-language speech sounds may be one source of individual variability in non-native speech sound learning. Regardless of whether native and non-native speech perception abilities are related, it is important to better understand the origins and potential consequences of individual differences in speech perception in native and second languages. A more complete understanding of this question may have an impact on our understanding of second language learning, speech perception in degraded listening conditions, or speech and language disorders.

Chapter 2

Replication and extension of overnight effects and structural neural correlates of non-native speech sound learning

2.1 Introduction

It is now well-established that adults have trouble learning non-native speech sounds and that individual vary a great deal in their ability to learn new speech contrasts (see Chapter 1 for a more in-depth discussion). Evidence of consolidation of non-native speech sounds and structural neural correlates of non-native speech sound learning has been found, but these findings are often inconsistent in the literature (see Chapter

1 and the following sections for more discussion). The goal of the current chapter is to replicate and extend these findings with a larger sample size, more modern statistical approaches, and surface-based MRI analyses.

2.1.1 Higher-powered replication of overnight effects in non-native speech sound learning

The idea that sleep helps learners consolidate learned information into long-term memory has a long history. In fact, consolidation was first discussed over a century ago when Müller and Pilzecker (1900) carried out their seminal studies on memory consolidation. However, the contributions of sleep to memory consolidation of non-native speech sounds have only recently received attention (see Earle & Myers, 2014 for review).

Only a handful of studies have tested overnight effects of consolidation of newly learned non-native speech sounds, and they often produce conflicting results. For example, Earle and Myers (2015b) trained two groups of participants to learn the voiced Hindi dental and retroflex stop consonants and tested their learning 12 and 24 hours later. One group was trained in the evening hours and one in the morning hours. They found that the evening group showed evidence of improvement after an interval of sleep, but the morning group did not. Qin and Zhang (2019) found similar effects in a non-native tone learning study: They found that evening-trained participants showed a trend toward overnight improvement, but the morning group got worse after the overnight interval.

Fuhrmeister, Smith, and Myers (2020) failed to replicate the finding that morning-

trained participants do not improve overnight. Although that study was testing a slightly different question, overnight improvement was seen in morning-trained participants for two different tasks measuring non-native speech sound learning. Some of the inconsistent findings in this area of research could however stem from differences in experiment design, specifically whether participants were exposed to phonetic variability coming from different talkers or phonological contexts. Borrowing from the visual perceptual learning literature, we have argued in previous papers (Fuhrmeister, 2019; Fuhrmeister et al., 2020) that exposure to variability during training may interfere with consolidation of new phonetic information because strongly learned information has been shown to be consolidated more effectively (e.g., Hauptmann, Reinhart, Brandt, & Karni, 2005; Shibata et al., 2017). Both Earle and Myers (2015b) and Qin and Zhang (2019) tested generalization to a new vowel context and talker, respectively. Fuhrmeister and Myers (2017) found that when simply testing a group of participants on generalization to an untrained vowel context, they did not show overnight improvement, while a group that was trained and tested on only one vowel context did improve after sleep. One reason the study by Fuhrmeister et al. (2020) may have shown overnight improvement is because participants were trained and tested on the sounds produced by one talker in one vowel context. Another study by Earle, Landi, and Myers (2017) found improvement on non-native speech sound learning tasks after sleep, but they also only trained and tested on one talker and one vowel context. Fuhrmeister and Myers (2020) trained and tested participants on two vowel contexts and found overnight improvement on the trained task (identification) but not on a discrimination task. Therefore, variability that is introduced even

by testing on another vowel context or another talker's voice may interfere with learning or consolidation processes. However, Earle and Myers (2015a) tested for generalization to a new talker and found overnight improvement, so it is still unclear whether we see inconsistent results in these types of studies because of phonological or talker variability.

Earle et al. (2017) additionally found that sleep duration predicted the amount of overnight improvement on non-native learning tasks. Specifically, they found that total sleep duration predicted overnight gains in discrimination performance and slow wave sleep predicted overnight gains in identification performance. Current work by Fuhrmeister and Fuchs (in preparation), however, measured total sleep duration and non-native speech sound learning and did not find that sleep predicted overnight improvement. Slight differences in design may explain this discrepancy; however, it is worthwhile to replicate this finding using a larger sample size, as both studies had small samples (Earle et al., 2017, $N = 17$; Fuhrmeister & Fuchs, in preparation, $N = 39$).

Another reason these studies may have found inconsistent results of overnight improvement is because, as discussed earlier, almost all published studies of non-native speech sound learning report a great deal of individual variability. Furthermore, many of these studies, including my own work, have had very small sample sizes and used statistical practices that are now out of date or that we now know to be problematic (e.g., using an ANOVA to analyze proportion correct data as in Fuhrmeister & Myers, 2017). Significant findings that result from small and noisy samples are very likely to be errors of magnitude (Type M errors)—effect sizes that are overexaggerated—or

errors of sign (Type S errors)—effect sizes that go in the opposite direction of the true effect (Gelman & Carlin, 2014; see also Button et al., 2013; Ioannidis, 2005, 2008). Therefore, it is likely that some of the significant findings that have emerged from this literature may be exaggerated effects or effects that even go in the wrong direction from the true effect. Finally, the field's overreliance on p-values as a threshold for publication further leads to findings that are not reproducible, as p-values and statistical significance itself is not very replicable (Amrhein et al., 2017; see also Gelman & Stern, 2006; Simmons, Nelson, & Simonsohn, 2011 for similar arguments). Although it is possible that the true effect of overnight improvement in non-native phonetic learning is small and is only detectable some of the time with the inherent large amount of variability in these samples, we simply do not have accurate estimates of the effect sizes in this literature, and this chapter takes a step toward remedying this situation.

The effects of sleep on non-native speech sound learning are certainly of theoretical interest, but also of practical interest. It is important to know if there are certain times of day for learning that could bolster consolidation (Earle & Myers, 2015b; Qin & Zhang, 2019), and it is important to know whether introducing variability can indeed have a detrimental effect on learning and consolidation (Fuhrmeister & Myers, 2017). Therefore, it is of interest to replicate these findings with a larger sample size. In the current study, we have collected a larger sample size than is typical for such studies (current study $N = 57$ for behavioral analyses, see section 2.1 below); other studies range from 13 to 35 per group). Furthermore, we are using more modern statistical approaches (e.g., mixed effects models) than several of the previous studies,

so we hope to offer a clearer picture of overnight effects in non-native speech sound learning and the influence of sleep.

In the current study, we first test whether we (conceptually) replicate previous studies' findings of overnight improvement on non-native learning tasks and whether sleep duration predicts overnight improvement.

2.1.2 Structural neural correlates of non-native speech sound learning

Analyses of brain structure have been used to better understand individual differences in non-native speech sound learning (see Chapter 1 and also (Golestani et al., 2007, 2002; Turker et al., 2017). As discussed in more detail in Chapter 1, MRI studies are particularly at risk for not being replicable (Button et al., 2013). First, they are expensive, so sample sizes will necessarily be smaller than many behavioral studies. Small samples combined with a large amount of individual variability lead to unreliable results, so it is important to replicate these findings with a larger sample size.

The current study expands on existing knowledge of structural relationships with non-native speech sound learning in three main ways. First, unlike previous studies, we have a measurement of discrimination of the non-native sounds that we obtained before and after training as well as after a period of offline consolidation. This allows us to look for structural relationships with naive discrimination ability (pretest scores), which typically predicts learning and retention of non-native speech contrasts (e.g., Fuhrmeister & Myers, 2020; Fuhrmeister et al., accepted). The fact that we

have behavioral measures after a period of offline consolidation allows us to test a new question, namely, whether hippocampal volume predicts improvement on the discrimination task after a period of sleep (see Earle & Myers, 2014, for a review on the contributions of sleep to non-native speech sound learning). The hippocampus has been found to play a role in sleep-related consolidation of newly formed memory traces, and an open question is whether hippocampal volume predicts improvement on the task after sleep on an individual level. Last, most studies examining the relationship between individual differences in brain structure and non-native speech sound learning have used volume-based approaches (e.g., voxel-based morphometry, Golestani et al., 2007, 2002; Turker et al., 2017; Wong et al., 2008, and few have used surface-based metrics (but see Rodriguez et al., 2018). Surface-based analyses have advantages over voxel-based morphometry because they allow analysis of several different structural metrics, such as surface area, cortical thickness, volume (which is the product of surface area and cortical thickness), curvature, and gyrification. There is evidence that cortical thickness and surface area result from different genetic processes (Wierenga, Langen, Oranje, & Durston, 2014; Winkler et al., 2010), so considering these metrics separately will give us more information about possible genetic differences that underlie non-native speech sound learning. Additionally, surface-based analysis allows more precise cortical parcellations than voxel-based approaches, especially in cortical areas that are highly gyrified because a single voxel sometimes encompasses cortical regions separated by a sulcus (e.g. Zatorre, Fields, & Johansen-Berg, 2012). Thus, the two regions of cortex may be functionally quite distinct but will be analyzed with one voxel. It is therefore of methodological interest

to conceptually replicate previous findings using surface-based analysis and to test whether cortical thickness or surface area are robust predictors of non-native phonetic learning.

As discussed in detail in Chapter 1, the following regions are known to be involved in perception and learning of native and non-native speech sounds: the pars opercularis region of the inferior frontal gyrus (Lee et al., 2012; Myers, 2007; Myers et al., 2009), the superior temporal gyrus (Chang et al., 2010; Myers, 2007), the transverse temporal gyrus and the planum temporale (Golestani et al., 2007, 2011; Turker et al., 2017; Wong et al., 2008), and the middle frontal gyrus (Luthra et al., 2019; Myers & Mesite, 2014; Myers & Swan, 2012). Therefore, we predict that we will see relationships between non-native learning tasks and structural measurements of surface area, cortical thickness, or volume in these regions. Of interest is whether surface area and cortical thickness of these areas will differentially predict behavioral performance.

Previous studies have found more instances of split or duplicate Heschl's gyri (i.e., early auditory areas) in faster learners of a non-native speech sound contrast (Golestani et al., 2007), expert phoneticians (Golestani et al., 2011), and individuals who were better at imitating non-native speech sounds (Turker et al., 2017). Taken together, these findings suggest that more gyrification (i.e., folding) in the transverse temporal gyri is related to superior speech perception abilities. However, other studies have found that more gyrification (split or duplicate gyri) in early auditory areas is related to phonological dyslexia (Leonard et al., 2001). These findings are limited, however, because many of these studies had small sample sizes (see discussion in

Chapter 1). The way gyrification has been measured in previous studies (counting the number of split or duplicate gyri) often results in comparisons between very uneven groups (e.g., comparing behavioral measures of subgroups of participants who had one vs. two gyri). Using surface-based metrics, we can obtain a continuous measure of gyrification, which will give us more statistical power (especially combined with our larger sample size, $N = 56$ for MRI analyses, see section 2.1 below). Based on previous findings, we expect the local gyrification of the transverse temporal gyrus to positively predict discrimination ability, possibly at all time points (i.e., before and after training).

2.1.3 Current study

The current study has three goals: First, we attempt to conceptually replicate previous findings showing improvement on non-native speech sound learning tasks after sleep, as well as whether sleep duration predicts the amount of overnight change using a larger sample size. Finally, we aim to extend previous findings on the structural neural correlates of non-native speech sound learning using surface-based analysis, a larger sample size, and a continuous measure of gyrification.

2.2 Method

2.2.1 Participants

Fifty-eight native speakers of English (43 female, 15 male) were recruited from the University of Connecticut community. Data from one participant was excluded from

all analyses reported in this chapter due to an equipment error. One participant's data was excluded from the MRI analyses because the participant did not complete that session of the experiment. Data from the remaining participants are reported below (behavioral analyses, $N = 57$; MRI analyses, $N = 56$). Participants reported having typical hearing and no history of speech or language disorders. We obtained informed consent from participants, following the guidelines of the University of Connecticut Institutional Review Board. Participants were compensated $10 per hour for behavioral tasks and $30 per hour for the MRI.

2.2.2 Stimuli and Materials

To assess non-native speech sound learning, participants in all experiments learned the voiced dental (/d̪/) and retroflex (/ɖ/) stop consonants found in Hindi (a difficult phonetic contrast for native English speakers to learn, e.g., Best et al., 2001). Stimuli for non-native speech sound learning tasks were recorded in a sound-attenuated booth by a female, native speaker of Hindi at the University of Connecticut in the Brain Imaging Research Center. Five recordings of each minimal pair nonword (/d̪ug/ and /ɖug/) were obtained. Stimuli were scaled to a mean amplitude of 65 dB SPL using Praat (Boersma & Weenink, 2013). All auditory stimuli were presented using over-ear headphones at a comfortable listening level that participants could adjust themselves. Visual stimuli consisted of "Fribbles," (novel objects that participants should have no familiarity with, stimulus images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, http://www.tarrlab.org/). All non-native phonetic learning tasks were

presented using OpenSesame experimental software (Mathôt, Schreij, & Theeuwes, 2012) on a desktop computer.

2.2.3 Procedure

Participants made a total of three visits to the lab: two behavioral sessions and one MRI session. The behavioral sessions were completed on two consecutive days. The first session took place between the hours of 5 and 9 PM, and the second session took place between 8 and 10 AM. In the first session, participants gave informed consent, then completed a phonetic training task to learn the Hindi sounds and were assessed on their identification and discrimination of the sounds. Sleep duration was also measured between the two behavioral sessions via an Actigraph GT3XP-BTLE wristwatch device (ActiGraph LLC, Pensacola, FL, USA). In the second session, participants were reassessed on their identification and discrimination of the Hindi sounds to measure retention and then completed two tasks to measure perception of native-language speech sounds. Some standardized cognitive and language tests were also administered, but those data are not reported here (see Appendix A for more information). The MRI session could take place at any time (before or after the behavioral sessions), as brain structure does not change rapidly as a result of phonetic training (see Golestani, 2014, for review). Structural MRI images were acquired from a 3-T Siemens Prisma with a 64-channel head coil. T1-weighted images were acquired sagitally using an MPRAGE sequence (TR = 2300 ms, TE = 2.98 ms, FOV = 256 mm, flip angle = 9 degrees, voxel size = 1 x 1 x 1 mm^3). Diffusion weighted images and magnetic resonance spectroscopy data were also collected but are not reported

here.

2.2.4 Non-native speech sound learning tasks

AX Discrimination. In order to assess pre- and post-training perceptual sensitivity to the Hindi sounds, participants completed an AX discrimination task. In this task, two of the minimal pair nonwords were presented auditorily (e.g., /d̪ug/ ... /ɖug/), and participants indicated whether they thought the words sounded the same or different. Participants completed 64 trials total with no feedback. For half of the trials, the initial speech sounds of each nonword came from the same speech category but were acoustically distinct recordings to discourage participants from using low-level detail of the acoustic signal to differentiate the sounds. Among the same trials, half of those consisted of two exemplars of the dental category, and half of the retroflex category. For the different trials, the onset speech sounds were from two different categories, and on half of these trials, the dental token was presented first, and for the other half, the retroflex was presented first.

Identification training and test. Immediately following the baseline AX discrimination measure, participants were familiarized with the nonwords that correspond to each novel visual stimulus. After that, participants completed 400 training trials with a two-minute break after the first 200 trials. On each training trial, participants saw two novel visual images on the screen and heard one word beginning with either the dental or retroflex sound presented over headphones. Minimal feedback was provided visually (e.g., “Correct!" or “Incorrect"). Identification tests consisted of 50 trials identical to training, except feedback was not provided.

2.2.5 Analysis approach

Behavioral tasks

Non-native speech sound learning tasks. For data from the discrimination tasks, d prime (d') scores were calculated [z(hits)-z(false alarms)] to account for response bias (Macmillan & Creelman, 2004). For identification assessments without feedback, we occasionally see that participants confuse the category labels even though they can differentiate the sounds, and this results in accuracy scores less than what would be expected by chance performance. We determined chance performance with a binomial test (less than 38% accuracy, $p < .05$), and participants' data whose total accuracy scores were less than 38% were recoded to reflect the label switching (i.e., 0 was recoded as 1 and 1 was recoded as 0).

To test whether participants improved overnight on the non-native learning tasks, we used mixed effects logistic regression models for the discrimination data and mixed effects logistic regression models for the identification data using the lme4 package (Bates et al., 2015) in R (R Core Team, 2008). Since we averaged over trials in the discrimination data (d' scores), we did not have enough data to estimate random slopes; therefore, we only included random intercepts for participant. In linear mixed effects models, p-values were estimated with the Satterthwaite method using the afex package (Singmann, Bolker, & Westfall, 2019). To determine the random effects structure of the model of identification data (where we had trial-level data), we used a backwards stepping procedure as in Matuschek, Kliegl, Vasishth, Baayen, and Bates (2017). All raw data and analysis scripts for behavioral analyses can be found at https://osf.io/cep8s.

MRI data

Structural MRI data were preprocessed with FreeSurfer's automated preprocessing pipeline (Dale, Fischl, & Sereno, 1999; Fischl, 2012). FreeSurfer reconstructs cortical surfaces into a two-dimensional triangular mesh and estimates the pial surface (boundary between gray matter and cerebral spinal fluid) and white matter surface (boundary between white matter and gray matter), from which surface area, cortical thickness, and volume can be calculated.

Whole-brain exploratory analyses

Many studies of non-native speech sound learning have not employed surface-based analysis, which can offer more information than voxel-based approaches. Therefore, we decided to do a whole-brain analysis in addition to region of interest analyses to explore whether non-native speech sound learning is predicted by clusters of surface area, cortical thickness, or volume that do not fall within defined regions of interest. A whole-brain analysis identifies clusters of vertices of structural metrics that differ as a function of group or are correlated with a continuous measure, as was done in the present study. To that end, we carried out a series of generalized linear models using the mri_glmfit command in Freesurfer. Separate analyses were carried out to test relationships between the measures of non-native learning (discrimination pretest and next-day posttest only to minimize the number of tests being done) and measures of surface area, cortical thickness, and volume, and for each hemisphere. Surfaces were smoothed with a Gaussian kernel with a full-width/half-max of 10mm. We used mri_glmfit-sim to implement a vertex-wise cluster forming threshold of .001

(Greve & Fischl, 2018) and a cluster-wise p threshold of .05 using non-directional tests. Bonferroni correction was applied to correct for tests over two hemispheres.

Region of interest analyses

For region of interest analyses, each vertex of the cortical surface is probabilistically assigned to a region according to an atlas. Regions of interest for the current study were selected from the Destrieux atlas in Freesurfer (Destrieux, Fischl, Dale, & Halgren, 2010). Based on previous literature, we identified the following bilateral regions of interest for our analyses of non-native measures: the pars opercularis region of the inferior frontal gyrus (Lee et al., 2012; Myers, 2007; Myers et al., 2009), the superior temporal gyrus (Myers, 2007), the transverse temporal gyrus and the planum temporale (Golestani et al., 2007, 2011; Turker et al., 2017; Wong et al., 2008), and the middle frontal gyrus (Luthra et al., 2019; Myers & Mesite, 2014; Myers & Swan, 2012). The FreeSurfer labels for these regions can be found in Table 4.1.

Table 2.1: Regions of interest and Freesurfer Destrieux atlas labels. All regions were tested bilaterally.

Region of interest	Destrieux atlas label
Middle frontal gyrus	G_front_middle
Inferior frontal gyrus (pars opercularis region)	G_front_inf-Opercular
Transverse temporal gyrus	G_temp_sup-G_T_transv
Planum temporale	G_temp_sup-Plan_tempo
Superior temporal gyrus	G_temp_sup-Lateral

To test whether structural measures of these regions were related to non-native speech sound learning, mixed effects models were fit to predict non-native discrimina-

tion[1] from structural metrics (surface area, cortical thickness, and volume) of each region of interest. Models were fit for each structural measurement separately. To account for differences in head size, total intracranial volume was added as a predictor to the models of surface area and volume. Cortical thickness is not as related to head size, so models were fit with measures of thickness as predictors without including total intracranial volume as a predictor. Details of each model can be found in the results section below.

We also tested whether hippocampal volume predicted overnight change in non-native discrimination or identification, as sleep-mediated memory consolidation may be beneficial for non-native learning (see Earle & Myers, 2014, for review). Subcortical structures require a volumetric segmentation procedure (described in detail in Fischl et al., 2002), so in contrast to cortical regions, we can only obtain volume measurements for subcortical structures. Note that the hippocampus is not listed in Table 4.1 because the volume measurements were derived from a different segmentation process, and these data were analyzed separately because we had specific a priori hypotheses about the contributions of hippocampal volume to *overnight change* in non-native speech sound learning tasks. Total intracranial volume was included as a fixed effect in this model to account for differences in head size.

[1]We originally planned to test relationships between structural measurements of regions of interest and non-native identification, but many of these models would not converge, likely because of the complexity due to the trial-by-trial data. We are working to find a solution for this.

Gyrification

We used Freesurfer to compute a local gyrification index using the -localGI flag in recon -all. The local gyrification index is the ratio of the smoothed pial surface to the cortical surface, and it is calculated at each vertex of the two-dimensional cortical surface (Schaer et al., 2012). The local gyrification index for a region of interest is the mean of the local gyrification indices at each vertex in each region of the cortical parcellation. Based on previous work, we were interested in the local gyrification of the bilateral transverse temporal gyri (Golestani et al., 2007, 2011; Turker et al., 2017; Wong et al., 2008).

2.3 Results

2.3.1 Non-native behavioral measures

Overnight improvement on non-native speech sound learning tasks

We first attempted to replicate effects of overnight improvement on non-native learning that some previous studies have reported (e.g., Earle & Myers, 2015b; Fuhrmeister et al., 2020). Data from all 57 participants went into these analyses.

Identification. To test whether participants improved overnight on the identification task (the trained task), we fit a mixed effects logistic regression model. The model predicted accuracy (0 or 1 for each trial), and time was included as a fixed effect (immediate posttest or next-day posttest), which was deviation coded (immediate posttest = -.5, next-day posttest = .5). The final model included random intercepts

for participant. The intercept of the model was significantly greater than zero, $\beta = 2.16$, $SE = .18$, $z = 11.85$, $p < .001$, indicating that participants' learning was above chance. There was no difference in performance in the two time points, $\beta = .05$, $SE = .08$, $z = .71$, $p = .48$, suggesting participants maintained training-induced gains after the overnight interval, but did not improve overnight (see Figure 2.1).

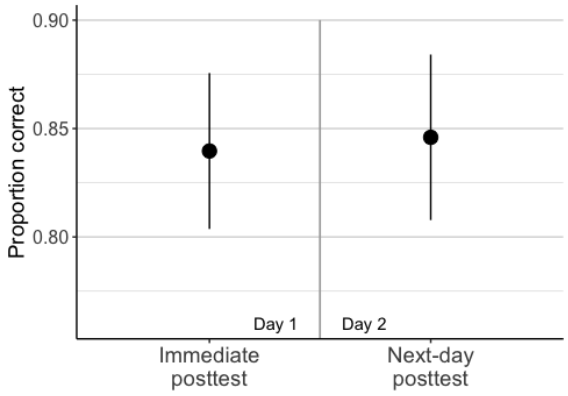


Figure 2.1: Non-native identification performance at each time point. Participants did not significantly improve after a period of offline consolidation. Error bars represent 95% confidence intervals.

Discrimination. To test for changes in discrimination performance over time, we fit a linear mixed effects model that predicted d' scores. Time was included as a fixed effect (pretest, immediate posttest, next-day posttest) and was backwards difference coded using the contr.sdif() function from the MASS package (Venables & Ripley, 2002) to test following contrasts: immediate posttest - pretest (improvement after training) and next-day posttest - immediate posttest (overnight improvement). Random intercepts for participant were included. The intercept of the model was significantly greater than zero, $\beta = 1.26$, $SE = .13$, $t = 9.79$, $p < .001$, indicating that

participants discriminated the sounds at above chance levels. There was a significant difference between the first two time points, $\beta = .72$, $SE = .11$, $t = 6.79$, $p < .001$, suggesting participants improved their discrimination as a result of training. However, the difference between the immediate posttest and the next-day posttest did not reach significance, $\beta = .18$, $SE = .11$, $t = 1.69$, $p = .09$, despite the numerical increase between these two time points (see Figure 2.2).

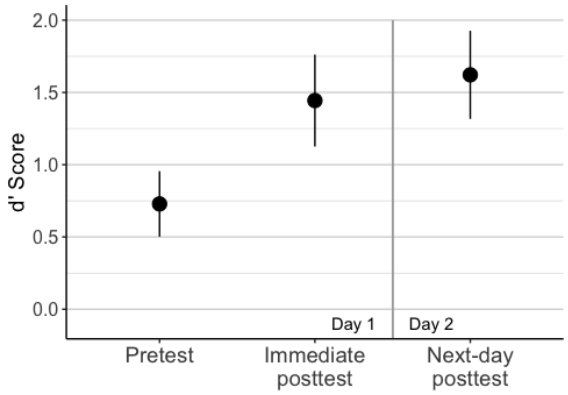


Figure 2.2: Non-native discrimination performance at each time point. Participants showed a significant increase in discrimination performance after training, but they did not significantly improve after a period of offline consolidation. Error bars represent 95% confidence intervals.

Sleep duration and overnight improvement

Although we did not see that participants as a group improved on the non-native tasks after an interval of sleep, we may see that an individual's sleep duration predicts overnight change. Sleep duration was measured in the current study with an Actigraph wristwatch device and also asked participants for a self-report of their sleep

duration to check the accuracy of the Actigraph data. Actigraph devices measure total sleep duration only. In the study by Earle et al. (2017), however, they measured sleep duration with a different device that measured individual sleep stages. One participant's sleep data was not recorded due to experimenter error, so the remaining 56 participants are included in the following analyses.

Identification. To test whether sleep duration in minutes predicted overnight improvement in non-native identification, we fit a mixed effects logistic regression model that predicted identification accuracy and included fixed factors of time (immediate posttest = -.5, next-day posttest = .5) and sleep minutes. For this and all other analyses of sleep duration, sleep minutes were corrected according to participant self-reports if self-reports of sleep onset or waking time deviated from the Actigraph data for more than 30 minutes. For example, if a participant reported going to bed at 10:30 PM and the sleep onset time from the Actigraph data was 10:45 PM, we assumed the Actigraph data was reliable and that the participant took 15 minutes to fall asleep. Corrected sleep minutes were then scaled and centered using the scale() function in R, which subtracts the column mean from each value and then divides it by the standard deviation of the observations. This was done to get the models to converge. Random intercepts for participant were included in the final model. The model revealed no difference between the two time points, no relationship between sleep duration and identification performance, and no interaction. The lack of interaction suggests that sleep duration did not predict overnight improvement on the identification task.

Discrimination. To probe the relationship between sleep duration and overnight

improvement in non-native discrimination (as was found in Earle et al., 2017), we fit a linear mixed effects model that predicted d' scores. Time was included as a fixed effect, but only the immediate posttest and next-day posttests were include because we were primarily interested in overnight change. Time was deviation coded (immediate posttest = -.5, next-day posttest = .5). Sleep duration in minutes (centered and scaled) was also included as a fixed effect, and random intercepts for participant were included. We found no difference between the two time points, no effect of sleep duration, and no interaction. The absence of an interaction again suggests that sleep duration did not have an effect on overnight improvement.

There is some debate about the best way to handle repeated measures in a pre-posttest design, specifically, whether it is most appropriate to test for an interaction with time as we did in the previous model or whether to include the baseline measure as a predictor in the model, similar to an ANCOVA approach (e.g. Van Breukelen, 2006). Because the main goal of this analysis was to replicate previous findings with a larger sample size, we wanted to ensure that our failure to replicate these findings was not due to our choice of analyses. Therefore, we fit a linear regression model that predicted d' scores on the next-day posttest. Predictors included the d' scores from the immediate posttest, scaled and centered sleep minutes, and their interaction. The overall fit of the model is as follows: $R^2 = .82$, $F(3,52) = 79.94$, $p < .001$. Unsurprisingly, the immediate posttest scores predicted the next-day posttest scores, $\beta = .85$, $SE = .06$, $t = 15.22$, $p < .001$, but sleep duration did not predict d' scores on the next-day posttest, nor did it interact with d' scores on the immediate posttest.

Finally, to do a closer replication of Earle et al. (2017), we fit one more linear regression model that predicted overnight change in d' scores (next-day posttest - immediate posttest) by sleep duration in minutes (scaled and centered). The fit of the model was not significant, $R^2 = .03$, $F(1,54) = 1.72$, $p = .20$, and sleep duration did not predict overnight change.

2.3.2 MRI analyses

Whole brain analyses

To explore whether structural measurements (surface area, cortical thickness, or volume) were related to non-native discrimination, we conducted a whole-brain analysis as described above. Full results can be found in the corresponding tables for each analysis. Anatomical regions for each cluster were determined by the cortical parcellations from the Desikan-Killiany atlas (Desikan et al., 2006).

Non-native discrimination pretest scores. First, we asked which structural variations predicted non-native discrimination at pretest, which can be seen as a measure of participants' naive sensitivity to the dental/retroflex contrast. We found negative relationships between non-native discrimination pretest scores and surface area in the left rostral middle frontal gyrus, left middle temporal gyrus, and left postcentral gyrus (see Table 2.2), as well as in the right fusiform gyrus (see Table 2.2). No relationships were found between the non-native discrimination pretest scores and cortical thickness. We found negative relationships between non-native discrimination pretest scores and volume in the left fusiform gyrus and the left posterior cingulate (see Table 2.3). Pretest scores were also negatively related to volume in the right

fusiform gyrus, right insula, and the right precuneus (Table 2.3, Figure 2.3).

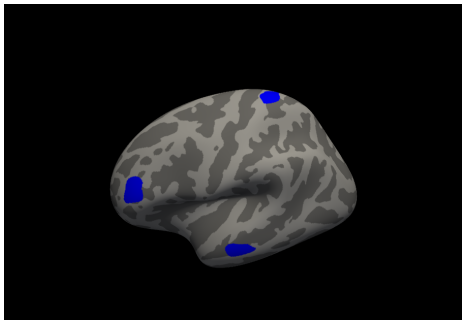
Table 2.2: Surface area and discrimination pretest. The r value is the average r values of the vertices in the cluster. Size, coordinates, cluster-wise p values, and anatomical region as defined from the Desikan-Killiany atlas are also indicated.

Cluster number	r-value	Size(mm^2)	MNIX	MNIY	MNIZ	p-value	Anatomical region
Left hemisphere							
1	-.45	502.61	-36.6	41.1	10.2	0.00679	rostral middle frontal gyrus
2	-.46	331.46	-55.8	-20.6	-15	0.04352	middle temporal gyrus
3	-.46	330.81	-20.7	-37.1	62.5	0.0443	postcentral gyrus
Right hemisphere							
1	-.47	726.87	37.4	-47.6	-19.6	0.0004	fusiform gyrus

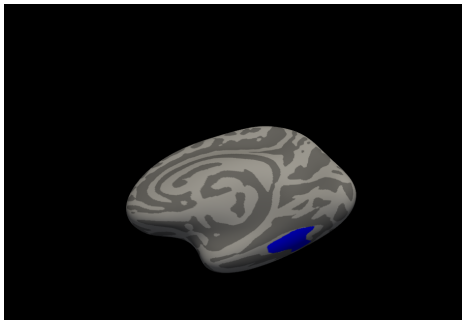
Table 2.3: Volume and discrimination pretest. The r value is the average r values of the vertices in the cluster. Size, coordinates, cluster-wise p values, and anatomical region as defined from the Desikan-Killiany atlas are also indicated.

Cluster number	r-value	Size(mm^3)	MNIX	MNIY	MNIZ	p-value	Anatomical region
Left hemisphere							
1	-.47	297.73	-38.4	-39	-22.5	0.0036	fusiform gyrus
2	-.46	201.24	-3.5	-13.4	34.3	0.02761	posterior cingulate
Right hemisphere							
1	-.48	754.15	31.4	-57.4	-15.1	0.0002	fusiform gyrus
2	-.44	264.72	38.3	-12.4	3.7	0.00639	insula
3	-.47	230.8	5.1	-66	31.4	0.01613	precuneus

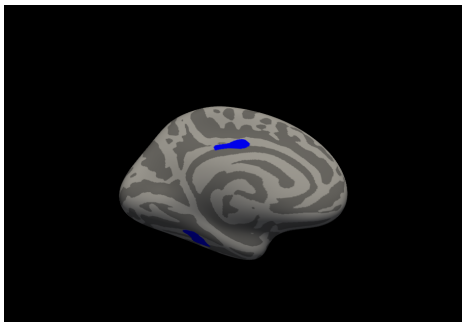
Non-native discrimination next-day posttest scores. Next, we asked whether structural variation predicted non-native discrimination ability at the next-day posttest. (Again, we did not test the immediate posttest to reduce the number of tests being carried out.) No relationships were found between non-native discrimination next-day posttest scores and measures of cortical surface area, cortical thickness, or cortical volume.



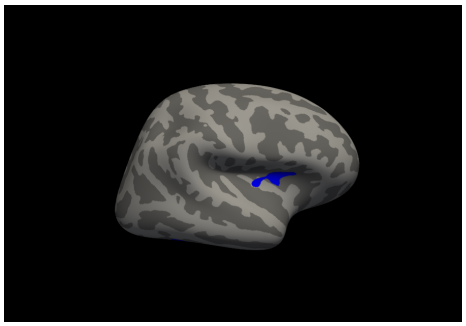
(a) Left hemisphere surface area and discrimination pretest.



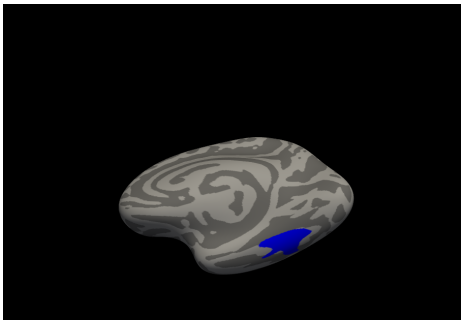
(b) Right hemisphere surface area and discrimination pretest.



(c) Left hemisphere volume and discrimination pretest.



(d) Right hemisphere volume and discrimination pretest: lateral view.



(e) Right hemisphere volume and discrimination pretest: medial view.

Figure 2.3: Clusters predicting discrimination pretest.

Region of interest analyses

To test whether surface area, cortical thickness, or volume of the pre-selected regions of interest predicted discrimination performance, we fit three linear mixed effects models

(one for each structural metric) that predicted d' scores. Time (pretest, immediate posttest, and next-day posttest) was included as a fixed effect, and it was treatment coded with pretest as the reference level. Structural metrics (surface area, cortical thickness, or volume) of each region of interest were included as predictors that were nested within the factor of time. This allowed us to test for simple effects of the relationship between the structural metric of interest (in each region of interest) and d' scores at each time point. Total intracranial volume was also included as a predictor in models testing surface area and volume. Random intercepts for participant were included, as well.

We only found that volume of the left inferior frontal gyrus (pars opercularis region) negatively predicted discrimination performance on the next-day posttest, when holding all other predictors constant, β = -2.43, SE = 1.21, t = -2.00, p = .049 (see Figure 2.4).

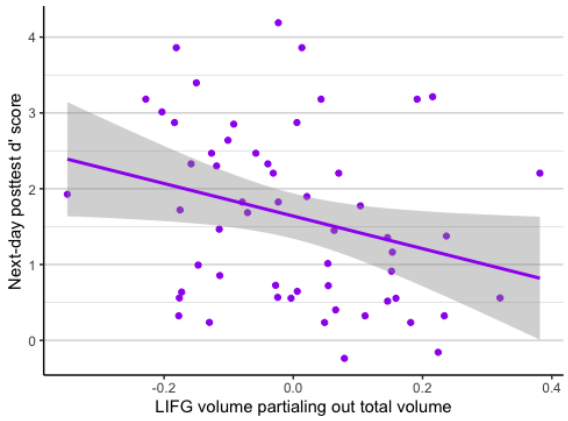


Figure 2.4: LIFG volume negatively predicts next-day posttest d' scores

Hippocampal volume

Discrimination. To test whether hippocampal volume predicted overnight change on non-native discrimination performance, we fit a linear mixed effects model that predicted d' scores. Fixed effects included time (deviation coded: immediate posttest = -.5, next-day posttest = .5), the interaction of hemisphere and hippocampal volume and their interactions (hemisphere was deviation coded: left = -.5, right = .5), and total intracranial volume to account for head size (lmer syntax: dprime ~ time*(hemisphere:hippocampal volume) + total intracranial volume). This allowed us to estimate a main effect of time, and simple effects of hippocampal volume on d' scores at each time point in each hemisphere independently without estimating a main effect for hemisphere. To facilitate model convergence, hippocampal volume and total intracranial volume were scaled. Random intercepts for participant were included in the model, as well. There was a difference in the two time points, β = -.90, SE = .39, t = -2.33, p = .02. Volume in either hemisphere did not predict d' scores, and neither did total intracranial volume. However, there was an interaction between hippocampal volume in the left hemisphere and time, β = 1.10, SE = .39, t = 2.79, p = .01, and an interaction between hippocampal volume in the right hemisphere and time, β = 1.07, SE = .38, t = 2.80, p = .01. This suggests the relationship between hippocampal volume in both hemispheres and d' scores was stronger at the next-day posttest than the immediate posttest. In other words, hippocampal volume in both hemispheres positively predicted overnight improvement on the discrimination task (see Figure 2.5).

Identification. To test whether hippocampal volume predicted overnight change on

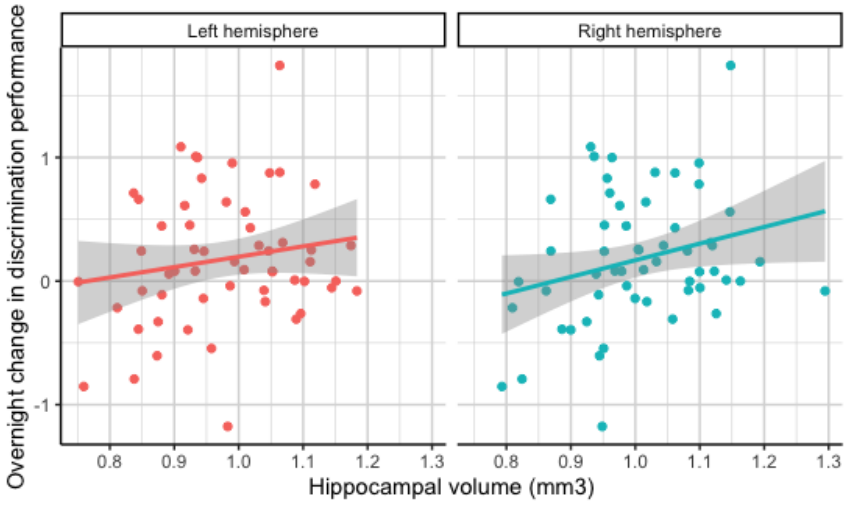


Figure 2.5: Volume of the left and right hippocampus positively predicts overnight change in discrimination performance. Note that overnight change is plotted as a difference score for ease of visualization.

the identification task, we fit a mixed effects logistic regression model that predicted identification accuracy (1 or 0). Fixed effects included time (deviation coded as in the discrimination analysis), the interaction of hippocampal volume and hemisphere (hemisphere deviation coded as before), and their interactions, and total intracranial volume was included as a fixed factor to account for head size (glmer syntax: accuracy ~ time*(hemisphere:hippocampal volume) + total intracranial volume). To facilitate model convergence, hippocampal volume and total intracranial volume were scaled. The random effects structure of the final model included by-participant random intercepts and slopes for time with correlation parameters set to zero. To get the model to converge, iterations were increased to 200,000 using the “bobyqa" optimizer in the glmerControl options. No significant predictors or interactions were found.

Gyrification

Discrimination. To test whether gyrification of the transverse temporal gyri predicted non-native discrimination measures, we fit a mixed effects model that predicted discrimination performance (d' scores). Fixed effects included time, the interaction of the local gyrification index of the transverse temporal gyrus and hemisphere, and their interactions (dprime ~ time*(hemisphere:local gyrification index)). Time was backwards difference coded using the contr.sdif() function from the MASS package (Venables & Ripley, 2002) to test the following contrasts: immediate posttest - pretest (learning) and next-day posttest - immediate posttest (retention). Hemisphere was deviation coded (left hemisphere = -.5, right hemisphere = .5). Random intercepts for participant were included, as well. We found a difference between the immediate posttest and the next-day posttest, β = -2.67, SE = 1.07, t = -2.49, p = .01. We additionally found an interaction of the local gyrification index in the left hemisphere and the difference between the immediate posttest and the next-day posttest, β = .59, SE = .22, t = 2.66, p = .008; and the same interaction in the right hemisphere, β = .58, SE = .22, t = 2.66, p = .008, suggesting that gyrification in the bilateral transverse temporal gyri was positively related to overnight change in discrimination performance (see Figure 2.6)[2]

[2]We tested these contrasts (learning and retention) based on previous findings by Golestani et al. (2007) who found that faster learners were more likely to have split or multiple transverse temporal gyri. However, after visually inspecting the data in Figure 2.6B, it looked the difference in slopes indicated by the interactions found in this analysis were a result of the negative relationship between gyrification and discrimination performance at the immediate posttest going away at the next-day posttest. We tested this by nesting the interaction between the local gyrification index and hemisphere within time to get simple effects of gyrification in each hemisphere at each time point. This model showed no significant relationships between gyrification and discrimination performance at any time point; however, the direction of the relationships between gyrification and discrimination performance were negative at the pretest and immediate posttest and close to zero at the next-day

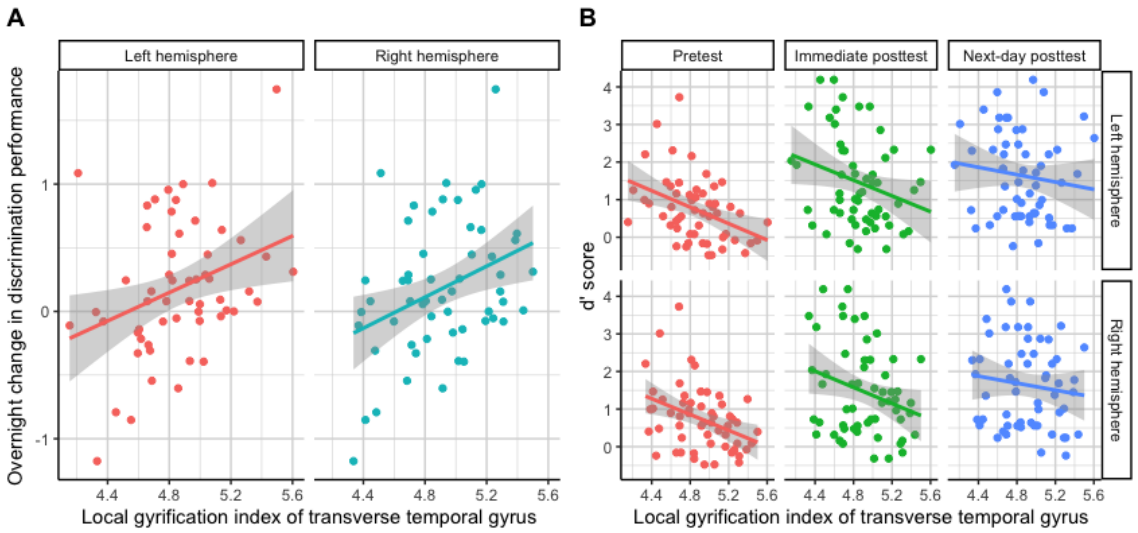


Figure 2.6: A. Local gyrification index of the bilateral transverse temporal gyri predicts overnight change in discrimination performance. B. Relationship between the local gyrification index of the transverse temporal gyri and d' scores at each time point.

Identification. To test whether gyrification measures of the transverse temporal gyri predicted non-native identification performance, we fit a mixed effects logistic regression model that predicted identification accuracy (1 or 0). Fixed factors included time (deviation coded: immediate posttest = -.5, next-day posttest = .5), the interaction of the local gyrification index of the transverse temporal gyrus and hemisphere (deviation coded as before), and their interactions (accuracy ~ time*(hemisphere:local gyrification index)). Random effects in the final model included by-participant random intercepts and slopes for time with correlations of random effects set to zero, and we used the glmerControl optimizer "bobyqa" to increase iterations to 200,000 to get the model to converge. The model revealed no significant effects or interactions,

posttest.

suggesting that gyrification of the transverse temporal gyri was not a predictor of identification performance.

2.4 Discussion

2.4.1 Conceptual replication of overnight effects and sleep duration in non-native speech sound learning

We first attempted to replicate previous findings (with a larger sample size) that participants improve on non-native learning tasks (identification and discrimination) after a period of sleep. We did not find any evidence of overnight improvement on the identification task, but we found a small, numerical increase in discrimination performance after sleep, but this increase did not reach significance. As discussed in the introduction, improvement after an overnight interval has only been found in a handful of studies, and in fact, in the seminal study testing the effects of sleep consolidation on non-native speech sound learning by Earle and Myers (2015b) did not even find improvement immediately after the overnight interval. Rather, they found improvement only when comparing performance from the evening-trained group on the immediate posttest to the posttest 24 hours after that (neglecting to do comparisons with the 12-hour posttest). Thus, even the first study to test this did not truly find overnight improvement. Improvement immediately after an interval of sleep has been found in a handful of studies (e.g., Earle et al., 2017; Fuhrmeister & Myers, 2017; Fuhrmeister et al., 2020), but most of the studies that have tested this have reported a great deal of individual variability among participants. Even when

testing this in the current study with a much larger than typical sample size ($N = 57$), we do not have strong evidence of overnight improvement.

One explanation for this is that the effect of overnight improvement is real but that it is a small effect and it will take a very large sample size to detect it. These results certainly point to the fact that as of yet, we (as a field) do not have accurate estimates of effect sizes for overnight improvement in these types of designs. Another explanation is that the studies that have found overnight improvement were based on small, noisy sample sizes, which will inherently overestimate significant findings (Gelman & Carlin, 2014). This is because the confidence intervals will be inherently large when there is a larger standard error due to variability in the data and a smaller sample size, so any significant differences that do emerge have to be larger in order for the confidence intervals not to overlap. An advantage we had in the current study is that we used more modern statistical approaches (mixed effects models) that have not been employed consistently in the literature. Combined with the larger sample size, the current study likely provides the best estimates of effect sizes that we currently have for overnight effects. However, we would recommend that future research run a large-scale, pre-registered replication study to obtain more reliable estimates of effect sizes. Given the large amounts of variability that are commonly observed in studies of non-native speech sound learning, the sample size needed to get reliable effect size estimates may be very large.

Next, we attempted to replicate the finding that sleep duration predicts overnight improvement. We could not do an exact replication of the Earle et al. (2017) findings, however, because we only had a measure of total sleep duration, whereas they

measured total sleep duration in addition to duration of several specific sleep stages. However, they found that total sleep duration predicted overnight improvement on the discrimination task, so we attempted to replicate that finding. We did not find that total sleep duration predicted overnight improvement on either the identification or discrimination tasks. One reason we may have failed to replicate this finding is because in Earle et al. (2017), they had measures of individual sleep stages, which were also entered into a regression model. From the current data set, we cannot comment on the replicability of the relationship between slow wave sleep duration predicting overnight change in the identification task; however, we find no evidence of a relationship between total sleep duration and overnight change in discrimination ability.

2.4.2 Structural relationships with non-native measures

In the current study, we tested whether individual variability in brain structure is related to behavioral performance on non-native speech sound learning tasks. The goal of this was to conceptually replicate and extend previous studies that have tested these relationships. In the current study, we used surface-based analysis to derive measures of different parts of brain structure separately, specifically surface area, cortical thickness, volume, and gyrification.

In a whole-brain analysis, we found several clusters where surface area or volume negatively predicted pretest discrimination, which can be though to reflect a listener's naive perception of the sounds or perhaps perceptual acuity. Negative relationships with pretest were found in the middle frontal gyrus and the insula. These findings are

somewhat consistent with fMRI studies showing inverse relationships with activation in frontal regions (including the insula) and non-native speech sound learning (e.g., Golestani & Zatorre, 2004; Myers & Swan, 2012). However, we did not see any relationships with discrimination performance *after* training in the whole-brain analysis. Although we typically see that pretest discrimination strongly predicts learning and retention (Fuhrmeister & Myers, 2020; Fuhrmeister et al., accepted), none of the relationships with pretest persisted after a delay. Having less surface area and volume of frontal regions may predict how well a listener can naively discriminate unfamiliar speech sounds, but it is unclear whether a reduction of surface area or volume in these regions has any long-term consequences for learning and retention of new speech sounds.

We also found negative relationships with discrimination pretest and brain regions that are not predicted by any theory or are not typically associated with speech perception (e.g., fusiform gyrus, posterior cingulate, precuneus). The fusiform gyrus has been implicated in second-language reading (Qu et al., 2019), and processing visual speech information shown from the face (Albonico & Barton, 2017). It is possible that the fusiform or other attention-related areas such as the cingulate are involved in speech processing; however, we leave it to future work to explicate the roles of these structures in speech perception.

In a region of interest analysis, we found that volume of the pars opercularis region of the left inferior frontal gyrus negatively predicts non-native discrimination after a period of offline consolidation. This finding is completely consistent with accounts that suggest that learners initially rely on frontal regions to perceive unfamiliar speech

sounds but that after time, they begin to rely more on posterior temporal regions (Myers, 2014, see also Chandrasekaran, Koslov, & Maddox, 2014; Chandrasekaran, Yi, & Maddox, 2014; Yi et al., 2016 for another theory that similarly predicts that reliance on frontal regions is suboptimal). This relationship emerged after a period of offline consolidation, so learners who were better able to discriminate the sounds after consolidation may be relying on other, more optimal regions such as primary or secondary auditory cortex, though we did not see any relationships with these regions of interest.

An interesting finding was that hippocampal volume positively predicted overnight change in discrimination performance. This is consistent with some findings in the memory consolidation literature, in which a larger hippocampus is related to better delayed or sometimes even immediate recall in patients with Alzheimer's disease (Köhler et al., 1998) or healthy young adults (Pohlack et al., 2014, but see meta analysis by Van Petten, 2004). The current findings may also lend support to the idea that learning non-native speech sounds is susceptible to domain-general memory processes and underscores the importance of these memory processes for learning new sounds. Learning non-native speech sounds is an interesting problem of learning and memory because of the difficulty and large amount of individual variability seen in adult learners, and the current results suggest that hippocampal volume may predict the degree to which a learner can take advantage of memory consolidation processes. Hippocampal volume has been found to predict other types of learning, such as statistical learning in children (Finn, Kharitonova, Holtby, & Sheridan, 2019) and it has been found to increase as a result of language learning (Bellander et al.,

2016; Mårtensson et al., 2012).

We additionally tested whether gyrification of the bilateral transverse temporal gyri predicted behavioral measures of non-native discrimination or identification. Based on previous findings (e.g., Golestani et al., 2007, 2001; Turker et al., 2017), we predicted that more gyrification of the transverse temporal gyrus in either hemisphere would predict better performance on non-native speech sound learning measures. We first found that a participant's local gyrification index was positively related to overnight change in non-native discrimination performance. We did not find any relationship with gyrification and identification performance.

At first glance, the relationship between gyrification and overnight improvement on the non-native discrimination task seems intuitive; however, the reason for this relationship was unexpected. Visual inspection of the plotted data in Figure 2.6B suggests that gyrification negatively predicted performance on the pretest and immediate posttest, but by the next-day posttest, the relationship was attenuated. This suggests that a greater amount of gyrification in these regions predicted poorer naive discrimination of the sounds, which does not seem entirely consistent with prior studies showing more instances of split or duplicate transverse temporal gyri in expert phoneticians (Golestani et al., 2011), faster learners of the Hindi dental/retroflex contrast (Golestani et al., 2007), or more accurate imitation (production) of the Hindi dental/retroflex sounds (Turker et al., 2017). Our findings seem to contradict previous findings; however, previous studies used different tasks and measures of gyrification. Specifically, all three of the previous studies mentioned manually identified the presence of split or duplicate transverse temporal gyri, and it is possible that the

local gyrification index used in the current study is capturing something different than the morphological differences observed in previous studies. Ultimately, future research will need to test whether a larger local gyrification index is related to split or duplicate gyri. Because these studies had smaller sample sizes to begin with (between 21 and 33 people), grouping participants based on whether they had single, split, or duplicate gyri results in very small sample sizes for some groups (sometimes less than 10 participants per group). These samples are likely not large enough to get reliable estimates of effects. The current study had a larger sample size (N = 56) and used a continuous measure of gyrification, so we did not have to split our sample into groups. Therefore, it is likely that we had more statistical power in the current study. It is also important to keep in mind that the Golestani et al. (2007) paper found more instances of split or duplicate gyri within faster learners as compared to slower learners, and the faster and slower groups of learners actually did not differ in their final posttest measure (an identification test). We would argue that a test of ultimate success after training is a more interesting or meaningful measure of learning than how quickly participants learned if they did not differ in the end. Golestani and colleagues (2007) did not find any relationships with gyrification and their final identification posttest, and our results from the current study are completely consistent with that.

The current finding that transverse temporal gyrification is related to overnight change in discrimination could be indicative of different learning strategies between good and poor perceivers. For example, the good perceivers may have been able to rely on auditory acuity throughout the entire training and testing process, whereas poor perceivers may rely more on consolidation processes to catch up after a period

of offline consolidation. In other words, the good perceivers may have been able to rely on acoustic differences of the stimuli both before and after training in order to discriminate them, while the poor perceivers may have tapped into category learning strategies which helped them more after a period of offline consolidation. It is also possible we are seeing hints of structural variation that is has been seen in phonological dyslexia as in the study by Leonard et al. (2001). In the end, the relationship between gyrification and non-native discrimination disappears after a delay, so it seems that any early disadvantages associated with more gyrification in these areas were not long-lasting.

2.5 Conclusions

The current study was a conceptual replication of some previous findings of overnight improvement with a larger sample size. We found no evidence of overnight improvement on non-native identification or discrimination (although there was a numerical trend in that direction for the discrimination task) and no evidence that sleep duration predicts overnight improvement. This suggests that overnight effects may be very small, and that future studies will need much better statistical power to detect them. We also extended previous findings of structural MRI correlates of non-native speech sound learning using surface-based analysis and found relationships between non-native discrimination and cortical surface area and volume in frontal regions, but no relationships between behavioral and cortical thickness.

Chapter 3

Behavioral relationships between native-language speech perception and non-native speech sound learning

3.1 Introduction

Many adults struggle to perceive and produce speech sounds in a second language (e.g., Bradlow et al., 1999), especially speech sounds that are perceptually similar to native-language sounds (e.g., Best & Tyler, 2007; Best et al., 2001). While theories of non-native speech sound learning are quite accurate in predicting which sounds will be difficult for second language learners to master, they do not account for the

individual variability among learners of the same language background. Despite the robust findings of individual differences in the non-native speech learning literature, the sources of this variability are poorly understood. One possible source of variability may be the native language. For instance, some evidence suggests there may be a common speech-specific mechanism that individuals rely on to process native and non-native speech sounds (Diaz et al., 2008). As will be described in detail in the following sections, theories of non-native speech sound learning predict that the perceptual proximity of non-native speech sounds to native speech categories is the source of difficulty for perceiving non-native speech sounds (e.g., Best & Tyler, 2007; Best et al., 2001; Kuhl, 1994; Kuhl et al., 2008). However, it is unknown whether these theories can be extended to predict which individuals will be more successful non-native speech sound learners. We test the hypothesis that individual differences in native-language speech category representations are one source of variability in learning new speech categories.

3.1.1 Individual differences in categorical perception of native-language speech sounds

Listeners perceive speech sounds categorically, i.e., differences between speech categories are highly detectable, while differences within speech categories are often poorly detected (e.g., Liberman et al., 1957, 1967). Since this seminal finding, much research in the field has expanded on this phenomenon but has given little attention to testing whether individuals differ systematically in whether their perception of phonetic categories is more categorical or more graded. This is an important area

of study because differences in sensitivity to within-category differences have been implicated in disorders, such as dyslexia (Serniclaes et al., 2004; Werker & Tees, 1987) and specific language impairment (Joanisse, Manis, Keating, & Seidenberg, 2000).

One potential limitation of previous studies is that classical, two-alternative forced choice identification tasks that are often used to measure categorical perception do not allow listeners to sufficiently demonstrate gradedness in their perception of speech sounds. For instance, when only given the choice between /d/ and /t/, many listeners who could otherwise detect subtle acoustic differences between two exemplars in the /d/ category with different voice onset time values, would still identify all or most /d/ tokens as belonging to the /d/ category because those tokens would still be better exemplars of /d/ than /t/ (see Chapter 1 for a more thorough discussion). Other tasks for measuring categorical perception offer the listener more opportunities to demonstrate graded perception of speech sounds. Some of these methods include eye-tracking (e.g., Clayards et al., 2008; McMurray et al., 2002), goodness judgments (Drouin et al., 2016; Miller, 1994), or visual analog scaling tasks, in which a listener moves a visual slider on a screen between two alternatives (Kapnoula et al., 2017; Kong & Edwards, 2016). Indeed, these studies suggest that listeners maintain sensitivity to the graded internal structure of speech categories (e.g., Clayards et al., 2008; Drouin et al., 2016; McMurray et al., 2002; Miller, 1994) and that considerable variability exists even among typically developing individuals in how graded or categorically speech categories are represented (Kapnoula et al., 2017; Kong & Edwards, 2016).

3.1.2 Theoretical predictions from non-native speech sound learning

There is no one-to-one correspondence between acoustic speech input and phonetic categories (Hillenbrand, Getty, Clark, & Wheeler, 1995; Peterson & Barney, 1952). Thus, listeners must efficiently map this variable input onto stable phonetic representations, and this mapping may result in categorical perception (i.e., good distinction of sounds that belong to different speech categories and poor distinction of different sounds from the same speech category). This may be advantageous for coping with variability in the speech signal. Although perceiving sounds categorically may be adaptive in the native language, it could pose a problem when the goal is to learn new speech sounds, especially when new speech sounds are perceptually similar to native-language categories whose representations are more stable. Several theories attribute difficulties in non-native speech sound learning to perceptual similarity of non-native sounds to native language speech categories (e.g., Best & Tyler, 2007; Flege, 1995; Francis & Nusbaum, 2002; Kuhl et al., 2008). For example, the voiced dental and retroflex stop consonants in Hindi are often difficult for native English speakers to learn because they are allophones of the alveolar /d/ category (e.g., in “width" or “address", Polka, 1991), and this makes it difficult to perceive them as separate sounds. While these models are highly accurate in predicting which non-native speech sounds will be most difficult to acquire for talkers of a given language background, they do not explain the individual variability that is so commonly observed. These models do not make explicit predictions about individual differences in non-native speech sound learning, but implicit in the models’ assumptions is that variation in

the perception of native-language speech category structure should predict how easy or difficult acquiring perceptual sensitivity to a new, non-native speech sound will be for an individual.

Perception-based models, such as the native language magnet model (Kuhl, 1994; Kuhl et al., 2008) and the perceptual assimilation model (Best et al., 2001; Best & Tyler, 2007) predict that non-native sounds that are perceptually similar to native-language sounds will be assimilated to existing native-language categories. In other words, speech sounds that are perceptually similar to native-language categories get perceived as exemplars of that sound. These models can generate predictions on an individual level: Both the perceptual assimilation model and the native language magnet model would predict a relationship between subtle variation in how native-language categories are perceived and how well someone can learn non-native speech sounds. For instance, if an individual showed perception of native-language categories that was more graded and less categorical (i.e., that person could distinguish subtle within-category differences), they may be less likely to assimilate similarly-sounding non-native speech sounds to existing native categories. In other words, the individual's native-language categories might be more flexible and therefore more easily allow for the creation of new categories. Further support for this notion comes from studies directly examining the developmental trajectory of speech categories. Burnham, et al. (1991) found that children are less categorical in their perception of stop consonants than adults. Because children typically achieve more native-like outcomes when learning non-native speech sounds in comparison to adults (e.g., Granena & Long, 2013; Stölten, Abrahamsson, & Hyltenstam, 2015), it is possible that more

graded perception of native-language speech sounds will help learners accommodate newly learned non-native speech sounds. Such a finding would be well supported by perception-based models.

Attention to dimension models (*attention-based models*, e.g., Francis & Nusbaum, 2002) would make the *opposite* prediction, namely that listeners who are more categorical in their native-language perception were better non- native speech sound learners. Attention to dimension models postulate that people learn to direct their attention to relevant acoustic cues in speech input, and the challenge in learning non-native speech sounds lies in learning to direct attention to previously un- or underattended cues. According to these models, the difference between good and poor learners of novel speech sounds should result from differences in directing attention to relevant cues, rather than perceptual abilities. This might suggest that those who show steeper categorization functions of their native-language sound categories can more efficiently direct their attention to relevant acoustic cues and ignore irrelevant ones. If these listeners perform better on non-native speech sound learning tasks, it would suggest that appropriate allocation of attention is a driving factor in individual differences in native and non-native speech processing. In that case, we might predict that those with steeper categorization functions would be better able to learn to direct their attention to different cues that distinguish new speech sounds, as measured by performance on non-native learning tasks. However, a finding in which graded perception predicts non-native speech sound learning would likely also be consistent with attention-based models because a “magnet-effect" could still be consistent with these predictions. For example, listeners may differ in

how flexibly they can attend to differences along a speech continuum. Although a finding in which more graded perception would make it difficult to adjudicate between attention and perception-based models, a finding in which more categorical perception predicted better non-native speech sound learning would be more easily explained by attention-based theories.

3.1.3 Current study

The aim of the current study is to test whether individual differences in categorical perception of native-language speech sounds predict success on a non-native speech sound learning task. We use a modified version of the visual analog scaling task used in Kapnoula et al. (2017) and Kong and Edwards (2016) to derive a continuous measure of how categorically or graded participants perceive native-language speech sounds, which we will refer to as categoricity. A finding in which categoricity negatively predicts non-native speech sound learning outcomes (i.e., participants who are less categorical are more successful on non-native tasks) would be consistent with perception-based models of non-native speech sound learning, such as the native-language magnet model (Kuhl, 1994; Kuhl et al., 2008) or perceptual assimilation model (Best et al., 2001; Best & Tyler, 2007). In contrast, a finding in which categoricity was positively associated with non-native speech sound learning would be more suggestive of attention-based models. Finally, no relationship between the two speech perception measures would lend support to the notion that non-native speech sound learning is a skill that is independent of native-language speech categoricity.

The discrete visual analog scale allowed us to measure not only categoricity, but

also how consistently participants rated the stimuli each time they were presented (i.e., how often they rated the first step on the continuum as “1" when it was presented). As discussed in Chapter 1, some recent evidence suggests that perception of native-language speech sounds becomes more graded throughout adolescence (McMurray et al., 2018) and that adult speech category representations are more graded than might be reflected by a two-alternative forced choice task (Kapnoula et al., 2017; McMurray et al., 2002, 2018). McMurray and colleagues (2018) argue that shallower categorization slopes in two-alternative forced choice phonetic identification tasks are a result of noisy representations, rather than the traditional view that they are due to more graded representations (e.g., Burnham et al., 1991). Therefore, we wanted to collect an exploratory measure of how noisy a participant's responses on the task were (response consistency). It is possible that more consistent responses on this task may predict non-native speech sound learning because those listeners have more precise representations of native-language speech sounds, which may be beneficial for learning new speech sounds.

3.2 Method

3.2.1 Participants

See Chapter 2 for participant information. Fifty-seven participants were included in the analyses reported in this chapter.

3.2.2 Stimuli and materials

See Chapter 2 for stimuli for non-native tasks.

To measure perception of native-language speech categories, participants rated tokens from two seven-step continua on a (modified) visual analog scale. One continuum consisted of a fricative contrast embedded in real words (sign-shine) and one consisted of a synthetic stop contrast of consonant-vowel syllables (ba-da). The ba-da continuum was created using a Klatt synthesizer at Haskins Laboratories. Stimuli for the sign-shine continuum were recorded by a female, native speaker of English, and waveform averaging in Praat (Boersma & Weenink, 2013) was used to create blends from 20% /s/ to 80% /s/ in 10% steps. While stop consonants are typically perceived categorically (e.g., Eimas, 1963), there is evidence that (at least some individuals) show more graded perception of fricatives (Healy & Repp, 1982; Repp, 1981). Therefore, we tested participants' perception of both a stop and fricative continuum.

Native-language visual analog scaling tasks were presented using E-Prime 3.0 (Psychology Software Tools, Pittsburgh, PA). Unlike in Kapnoula et al. (2017) and Kong and Edwards (2016), the visual analog scale used in the current study had seven discrete points along the line. With this discrete version of the scale, listeners had one response option for every point on the continuum that they heard during the task. This allowed us to measure not only how categorical a participant's responses were, but we also were able to obtain a measure of how consistent a participant's responses were for repeated tokens on the continuum (i.e., whether they responded with the same point on the continuum each time a specific token was played).

3.2.3 Procedure

Participants came to the lab for a total of two sessions. The first session took place in the evening hours (between 5 and 9 PM) and included consent and non-native phonetic training and assessments. The second session took place the following morning between 8 and 10 AM. This session included a test of retention of non-native speech sounds, the visual analog scaling task to measure categorical perception of native-language speech sounds, and a selection of standardized cognitive tests.

Non-native speech sound learning tasks. See Chapter 2 for a description of non-native speech sound learning tasks.

Native-language speech perception measures. Participants completed two tasks measuring categoricity (or gradedness) of native-language speech perception. In each task, auditory tokens taken from one of two continua were presented (/sa͡ɪn-ʃa͡ɪn/) or (/bɑ-dɑ/), and the order was counterbalanced. Participants indicated their response on a modified version of a visual analog scale. In this task, a line with discrete tick marks appeared on a screen and participants moved a slider to different points on the line between two stimuli (in this case, words/syllables) to indicate how, for example, sign-like or shine-like each stimulus token sounded (see Figure 3.1). Because we are interested in individual differences in how categorically a listener perceives native-language sounds, we wanted participants in this study to have more graded response options.

Standardized cognitive tests. Because individual differences in language ability and cognitive skills may influence native and non-native speech processing, we administered two tests of native-language phonological processing: the non-word repetition task

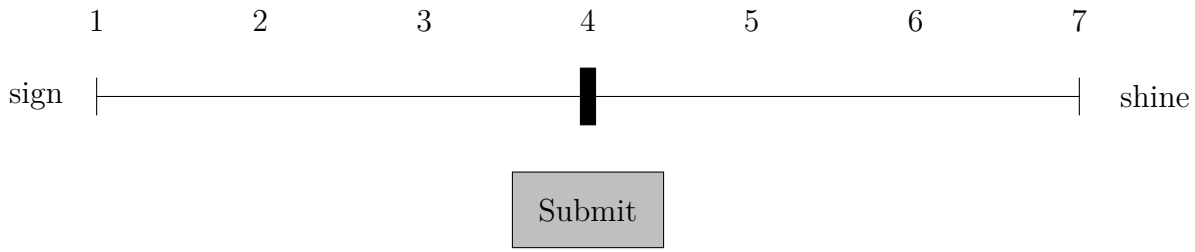


Figure 3.1: Sample trial of the visual analog scaling task.

from the Comprehensive Test of Phonological Processing (CTOPP, Wagner, Torgesen, Rashotte, & Pearson, 1999) and the sound blending task from the Woodcock-Johnson III (WJ-III, Woodcock, McGrew, Mather, et al., 2001), and two tests of working memory: auditory working memory, and numbers reversed tasks from the WJ-III (Woodcock et al., 2001). Those data are not reported here, but more information can be found in Appendix A.

3.2.4 Analysis approach

Non-native speech sound learning tasks. See Chapter 2 for analysis approach for non-native speech sound learning tasks.

Native-language speech perception measures. To obtain a measure of how categorically or graded an individual perceives native-language speech sounds, we ran a mixed effects non-linear regression model that fit responses to a 3-parameter logistic function (3-parameter because the 4-parameter model never converged) for data from the /bɑ-dɑ/ continuum. For the /s͡aɪn-ʃ͡aɪn/ continuum, we fit a mixed effects non-linear 2-parameter logistic model because the 3-parameter model did not converge. These models were run in R (R Core Development Team, 2008) using the

nlme package (Pinheiro, Bates, DebRoy, & Sarkar, 2019). The 3-parameter model estimates coefficients for the maximum asymptote, the inflection point (conceptually understood here as the category boundary), and the slope of the function (higher slope values indicate more categorical responses). The 2-parameter model estimates the inflection point and slope. A measure of response consistency for each participant was obtained by taking the mean of the residuals from each model for each participant and squaring them to avoid negative values. This means that larger values represented *less* consistent responses because they were derived from the residuals. To make interpretation of the results more intuitive, however, we changed the sign of the response consistency measure so that larger values would represent *more* consistent responses on the task. The measures of slope and consistency were entered into further analyses described below. Descriptive statistics on these measures are included in Table 3.1. See also Figures 3.2 and 3.3. All raw data and analysis scripts can be found at https://osf.io/cep8s.

Slope mean	Slope SD	Slope min.	Slope max.	Consistency mean	Consistency SD	Consistency min.	Consistency max.
/bɑ-dɑ/ continuum							
.57	.31	.17	1.95	-1.11	.39	-2.21	-.28
/sa͡m-ʃa͡m/ continuum							
.08	.02	.04	.11	-1.11	.39	-1.99	-.24

Table 3.1: Descriptive statistics for categoricity and consistency measures for each continuum ($N = 57$).

3.3 Results

For each analysis described here, we first included our measure of categoricity (or slope) in the model and time point as fixed effects, as this was our hypothesis-driven

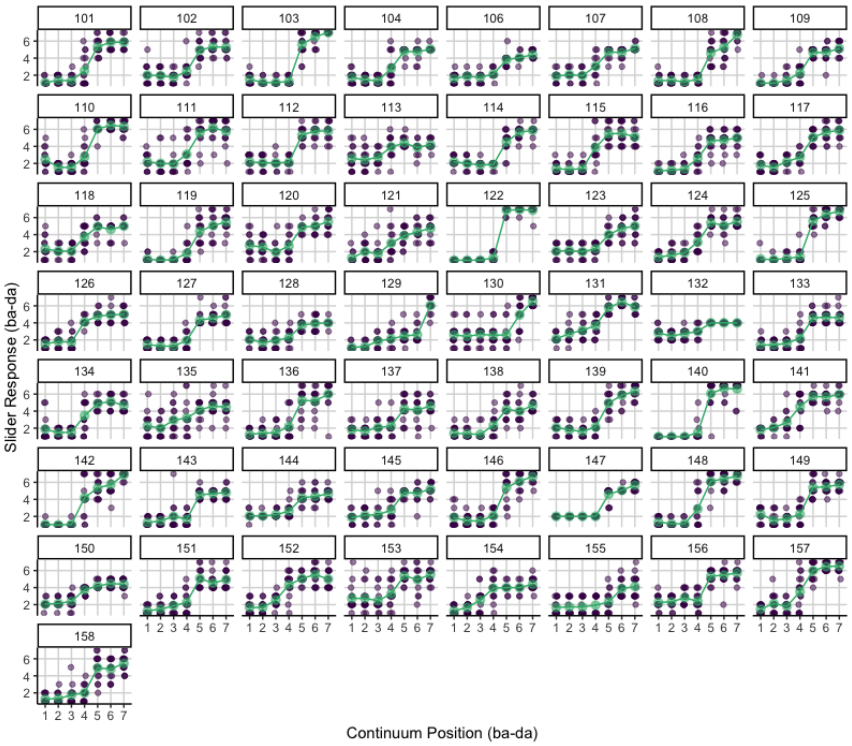


Figure 3.2: Individual plots of responses on the /bɑ-dɑ/ continuum. Purple circles represent responses on individual trials (darker purple means that there were multiple responses of that value for that point on the continuum), and green circles represent the mean of the responses for that point on the continuum. Error bars denote standard error.

analysis. The measure of response consistency was exploratory, so that measure was added to the original models after first assessing the effect of categoricity on non-native speech sound learning[1]. Response consistency and categoricity measures were not correlated. There were participants that looked like potential outliers for the measure of categoricity from the ba-da continuum, so we fit the statistical models

[1]We additionally fit exploratory models predicting discrimination and identification performance with fixed effects of time, consistency, and their interaction. No significant effects or interactions were found.

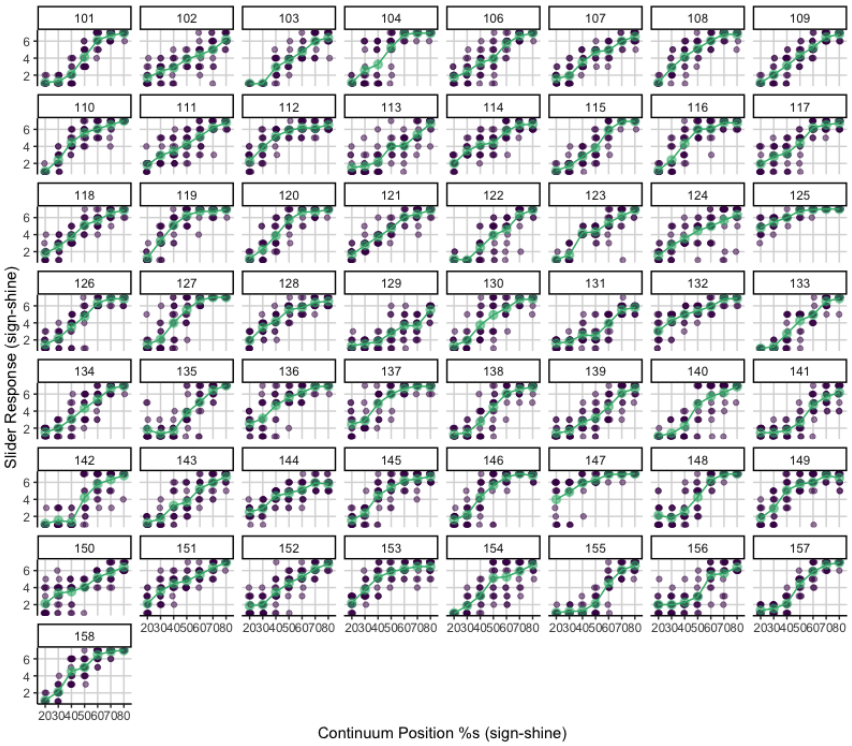


Figure 3.3: Individual plots of responses on the /saɪn-ʃaɪn/ continuum. Purple circles represent responses on individual trials, and green circles represent the mean of the responses for that point on the continuum. Error bars denote standard error.

with and without those participants' data. The pattern of results did not change, so results are only reported with the full sample.

3.3.1 Discrimination performance

Categoricity and consistency measures from the ba-da continuum. To test whether categoricity measured by the ba-da continuum predicted discrimination performance, we fit a linear mixed effects model that predicted d' scores. Fixed effects included time (pretest, immediate posttest, and next-day posttest), which was backwards difference

coded using the contr.sdif() function from the MASS package (Venables & Ripley, 2002) to test contrasts of next-day posttest-immediate posttest and immediate posttest-pretest, and the slope coefficients from the ba-da categorization task (categoricity). Random effects included by-participant random intercepts. The model revealed only a non-significant marginal difference between the pretest and immediate posttest, $\beta = .52$, $SE = .29$, $t = 1.79$, $p = .08$. Categoricity did not predict discrimination performance nor did it interact with time.

To test whether within-participant response consistency on the ba-da categorization task predicted discrimination performance, we added the measure of response consistency to the previous model. None of the measures were significant predictors of discrimination, and there were no interactions.

Categoricity and consistency measures from the sign-shine continuum. To test whether categoricity measured by the sign-shine continuum predicted discrimination performance, we fit another linear mixed effects model that predicted d' scores and included fixed effects of time (coded as before) and the slope coefficients from the sign-shine categorization task. Random effects included by-participant random intercepts. This model revealed a difference between the pretest and immediate posttest, $\beta = 1.13$, $SE = .55$, $t = 2.10$, $p = .04$, but the categoricity measure did not predict discrimination performance, and there were no interactions.

We again wanted to test whether response consistency on the sign-shine task predicted discrimination performance, so we added that measure to the previous model, and results were similar. We observed a difference between the pretest and immediate posttest, $\beta = 3.61$, $SE = 1.72$, $t = 2.10$, $p = .04$, but no other significant

effects.

3.3.2 Identification performance

Categoricity and consistency measures from the ba-da continuum. To test whether categoricity as measured by the ba-da task predicted performance on the trained task (identification), we fit a mixed effects logistic regression model that predicted accuracy on the identification task (0 or 1). Fixed effects included time (immediate posttest and next-day posttest), which was deviation coded (immediate posttest = -.5, next-day posttest = .5) and the ba-da categorization slope coefficient (categoricity). The final model included random intercepts for participant. This model revealed a difference between the two time points, β = .46, SE = .16, z = 2.91, p = .004, no main effect of categoricity in the ba-da task, and an interaction between time and categoricity, β = -.73, SE = .25, z = -2.92, p = .003. Based on the way the factor of time was coded, this might suggest that the categoricity measure was a stronger predictor of identification at the immediate posttest than the next-day posttest. To unpack the interaction, we fit a model that predicted identification accuracy and nested categoricity within the fixed effect of time (also with random intercepts for participant). Nesting fixed effects allows us to test the simple effects of a factor at each level of another factor (Schad, Vasishth, Hohenstein, & Kliegl, 2020)—in this case, the relationship between categoricity and identification performance at each time point separately without estimating the main effect of categoricity. This model again indicated a difference between the two time points, β = .46, SE = .16, z = 2.91, p = .004, and categoricity did not significantly predict identification performance at

either time point: immediate posttest, β = .57, *SE* = .60, z = .96, p = .34; next-day posttest, β = -.16, *SE* = .59, z = -.27, p = .79. This suggests that the interaction was a result of the different signs of the slope coefficients: The slopes of categoricity predicting identification accuracy differed from each other at each time point, though neither was a significant predictor of identification performance.

To test whether response consistency in the ba-da task predicted identification performance, we fit another mixed effects logistic regression model that predicted accuracy and included fixed effects of time, categoricity, consistency, and their interactions. This final model also included random intercepts for participant. The model revealed no significant predictors or interactions.

Categoricity and consistency measures from the sign-shine continuum. To test whether categoricity as measured by the sign-shine continuum predicted performance on the identification task, we fit a mixed effects logistic regression model that predicted accuracy. Fixed effects included time (immediate posttest and next-day posttest), which was deviation coded as before, the sign-shine categorization slope coefficient, and their interaction. The final model included random intercepts for participant. This model revealed no difference between the time points and no interactions.

To test whether response consistency on the sign-shine continuum predicted identification accuracy, we added this measure to the previous model. This model included by-participant random intercepts, and to get the model to converge, we used the optimizer “bobyqa" in the glmerControl options and increased the iterations to 200,000. We found a difference between the two time points, β = -1.86, *SE* = .69, z = -2.68, p = .01; an interaction between time and the sign-shine slope coefficient, β =

22.94, SE = 8.34, z = 2.75, p = .006; an interaction between time and the consistency measure, β = -1.90, SE = .62, z = -3.08, p = .002; and an interaction among time, the sign-shine slope, and consistency, β = 22.86, SE = 7.31, z = 3.13, p = .002.

To unpack these interactions, we fit a nested model. We originally tested the full model with interactions because we did not have any a priori assumptions about whether categoricity or response consistency would show a different relationship with identification accuracy at the immediate posttest or the next-day posttest (before or after a period of offline consolidation). To unpack the interactions, we fit a model that again predicted identification accuracy. The fixed effect structure nested categoricity (the slope coefficient of the sign-shine task), response consistency, and their interaction within time, and random effects included by-participant random intercepts. We again used the optimizer “bobyqa" in the glmerControl options and increased the iterations to 200,000 to get the model to converge. The model revealed a difference in the two time points, β = -1.86, SE = .90, z = -2.08, p = .04. In addition, there was an interaction between categoricity and response consistency at the next-day posttest, β = 23.03, SE = 8.61, z = 2.68, p = .007. Interactions with two continuous variables are difficult to interpret, so for ease of visualization, Figure 3.4 shows categoricity as a median split (high and low) and shows the relationship between consistency and identification accuracy for higher categoricity values (more categorical) and lower categoricity values (more graded). From the plot it appears that for more categorical responders, there was a positive relationship between consistency and identification accuracy, but for more graded responses, there was no relationship between consistency and identification accuracy. In other words, individuals who

responded more categorically and more consistently on the fricative categorization task were more accurate on the non-native identification task after a period of offline consolidation.

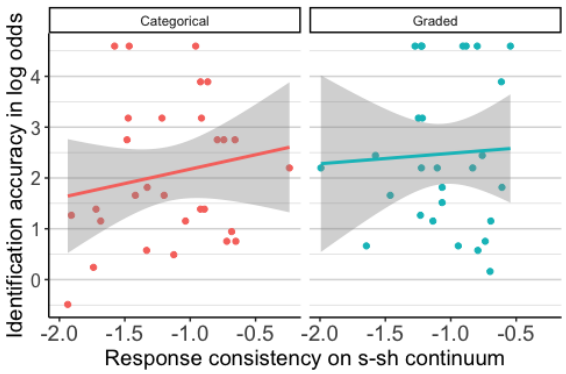


Figure 3.4: Next-day identification posttest data with categoricity shown as a discrete variable (median split: high and low) for visualization purposes only. Participants with more categorical and more consistent responses were more accurate on the identification task on the next-day posttest.

3.4 Discussion

Prominent theories of non-native speech sound learning could potentially be extended to predict that an individual who is able to perceive speech sounds in a more graded manner would be less likely to assimilate non-native tokens to native-language speech categories. This should result in superior learning or retention of non-native speech sounds. In the current study, we tested whether a measure of how categorically an individual perceives native-language speech sounds (from both a stop and fricative continuum) or whether a measure of how consistently participants responded to the

stimuli predicted their discrimination or identification of non-native speech sounds. We did not find that individual differences in categorical perception or response consistency from either the stop or fricative continuum predicted non-native discrimination or identification performance. We did, however, see that more categorical combined with more consistent responses on the fricative continuum (s-sh) predicted identification accuracy, but only on the second day, after a period of offline consolidation.

In general, the current findings do not lend support to the idea that categoricity of native-language speech sounds is what makes non-native speech sound learning hard at the individual level. It is intuitive to assume that perceptual reorganization in infancy or early childhood (e.g., Kuhl et al., 2006; Werker & Tees, 1984) is related to the development of categorical perception. Specifically, if we lose perceptual sensitivity to certain non-native speech sounds and those non-native speech sounds are then assimilated to native-language speech categories during perception (e.g., Best et al., 2001), it is logical to assume that categorical perception is what is responsible for assimilating perceptually similar non-native sounds to native-language categories. However, these processes may not be as related as is commonly assumed. Another possibility is that theories of non-native speech sound learning cannot be extended to make predictions at the individual level. There is ample experimental evidence that some phonetic contrasts are more difficult for learners of a certain native-language background to perceptually disambiguate (e.g., Best et al., 1998) and that the way non-native sounds map onto native-language sounds affects perception (e.g., Best et al., 2001). Although these theories are fairly accurate at predicting which sounds will be most difficult to learn for a learner with a certain native language, we currently

do not (to our knowledge) have a theory that predicts which individuals will be most successful at learning to perceive difficult non-native speech sound contrasts.

It is also possible that our behavioral measure of categoricity in the present study was not optimal for measuring individual variability. Although this and similar types of visual analog scaling tasks are certainly a more sensitive behavioral measure than a two-alternative forced choice task (e.g., Kapnoula et al., 2017), it may not have been sensitive enough. For example, we used discrete points on the visual scale, whereas previous studies have used continuous scales. Perhaps if we had used a more sensitive measure such as a traditional visual analog scaling task or an online measure such as eye tracking to obtain a measure of categoricity, we would have found a relationship between categoricity and non-native speech sound learning (e.g., McMurray et al., 2018). In short, there are a variety of tasks that have been used to measure categorical perception, and it is possible that the measure we chose was not quite sensitive enough to capture enough variability to measure individual differences in categorical perception.

The finding that more categorical combined with more consistent responses on the s-sh continuum predicted identification accuracy after a period of offline consolidation is puzzling. Some evidence suggests different patterns of categorical perception for different types of sounds (fricatives, stops, vowels; e.g., Eimas, 1963; Healy & Repp, 1982; Repp, 1981), which may be why we only saw this relationships in the s-sh continuum. Listeners who were more categorical and also more consistent in their responses may simply be good phonetic categorizers; they show both reliable behavior on the task as well as a sharp delineation between categories, and this could have

helped them succeed on the category learning (identification) task. This may be why this relationship emerged in the non-native identification task, rather than the discrimination task, which may rely on different perceptual (Guenther, Husain, Cohen, & Shinn-Cunningham, 1999) or memory mechanisms (see Earle & Myers, 2014, for review).

3.5 Conclusion

Current theories of non-native speech sound learning predict that native language speech categories are the source of difficulty for perceptual learning of non-native speech sounds. The results from the current study pose a challenge for the field because we did not find that individual differences in native-language phonetic representations (i.e., individual differences in categorical perception) predicted non-native speech sound learning, at least with the tasks used here. This suggests that non-native speech sound learning may be more separate from or at least less dependent on native-language speech perception than previously thought and that perhaps other cognitive skills underlie non-native speech sound learning. Furthermore, the question of why individuals vary so much in their ability to learn non-native speech sounds remains open.

Chapter 4

Structural neural correlates of categorical perception

4.1 Introduction

In this chapter, our goal is to establish whether certain measurements of brain structure (surface area, cortical thickness, volume, or gyrification) predict individual differences in categorical perception and response consistency on the visual analog scaling task described in Chapter 3. To our knowledge, this is the first study to test relationships between brain structure and individual differences in categorical perception of native-language speech sounds. This is of interest because how categorically or graded an individual perceives speech sounds has been found to be related to language and reading disorders (e.g., Serniclaes et al., 2004; Werker & Tees, 1987), and brain structure can often suggest whether abilities are learned or innate due to the

developmental trajectory of different aspects of brain structure (e.g., gyrification patterns, cortical thickness). Because (to our knowledge) no previous studies have tested which regions' structural metrics predict individual differences in categorical perception of speech sounds, we rely on the functional MRI literature to make predictions.

As reviewed in the introduction chapter, the *functional* MRI literature suggests that the following regions are involved in native-language categorical perception: frontal regions, including the inferior and middle frontal gyri (Blumstein, Myers, & Rissman, 2005; Lee et al., 2012; Myers, 2007; Myers et al., 2009; Myers & Mesite, 2014; Myers & Swan, 2012; see also Golestani et al., 2011, for a structural MRI study about phonetic expertise), the superior temporal gyrus (Bidelman et al., 2013; Blumstein et al., 2005; Chang et al., 2010; Desai, et al., 2008; Myers, 2007), the planum temporale (Schremm et al., 2018), and the transverse temporal gyri (Golestani et al., 2007, 2011; Turker et al., 2017). Frontal regions have been shown to be sensitive to category boundaries or phonetic competition (e.g., Blumstein et al., 2005; Myers, 2007; Myers et al., 2009) or show categorical-like responses (e.g., more sensitivity to between-category changes than within-category changes to stimuli; Luthra et al., 2019; Myers & Mesite, 2014, whereas temporal regions show sensitivity to the internal category structure of phonemes (Blumstein et al., 2005; Myers, 2007). Therefore, we expect to find relationships with structural measures from these regions and individual measures of categoricity and response consistency. More specifically, we predict that we will find relationships with brain structure and gradiency of perception in auditory/temporal regions, but individuals who are more categorical will show

structural differences in frontal regions, such as the inferior frontal gyrus or middle frontal gyrus.

4.2 Method

4.2.1 Participants

See Chapter 2 for a description of participants. Data from the same 56 participants in Chapter 2 reported in the MRI analyses are reported in this chapter.

4.2.2 Stimuli and Materials

See Chapter 2 for stimuli and materials for behavioral tasks.

4.2.3 Procedure

See Chapter 2 for procedure and MRI data acquisition.

4.2.4 Native-language speech perception measures

See Chapter 3 for native-language speech perception measures.

4.2.5 Analysis approach

See Chapters 2-3 for descriptions of analysis approaches for behavioral data, MRI preprocessing, and region of interest analyses. For reference, the regions of interest

and their Destrieux atlas labels (Destrieux et al., 2010) are included again here in Table 4.1.

Whole-brain exploratory analyses

We know of no studies that have tested the relationship between individual differences in brain structure and categorical perception of native-language speech sounds, so it is of interest to conduct an exploratory whole-brain analyses in addition to hypothesis-driven regions of interest analyses to identify possible differences in brain structure that may not be predicted by the literature. In addition, the regions of interest that we chose are large, and it is possible that only smaller clusters within these regions predict categoricity. As in Chapter 2, we did separate analyses using mri_glmfit for surface area, cortical thickness, and volume, and for each hemisphere to measure relationships between structure and native-language perception (categoricity and consistency). Surfaces were smoothed with a Gaussian kernel with a full-width/half-max of 10mm. We used mri_glmfit-sim to implement a vertex-wise cluster forming threshold of .001 (Greve & Fischl, 2018) and a cluster-wise p threshold of .05 using non-directional tests. Bonferroni correction was applied to correct for tests over two hemispheres.

Gyrification

Local gyrification index for the bilateral transverse temporal gyri was calculated as before in Chapter 2.

Table 4.1: Regions of interest and Freesurfer Destrieux atlas labels. All regions were tested bilaterally.

Region of interest	Destrieux atlas label
Middle frontal gyrus	G_front_middle
Inferior frontal gyrus (pars opercularis region)	G_front_inf-Opercular
Transverse temporal gyrus	G_temp_sup-G_T_transv
Planum temporale	G_temp_sup-Plan_tempo
Superior temporal gyrus	G_temp_sup-Lateral

4.3 Results

4.3.1 Whole brain analyses

Full results can be found in the corresponding tables for each analysis. Anatomical regions for each cluster were determined by the cortical parcellations from the Desikan-Killiany atlas (Desikan et al., 2006).

4.3.2 Native-language categoricity

No relationships were found between native-language categoricity (for either continuum) and measures of cortical surface area, cortical thickness, or cortical volume.

4.3.3 Native-language consistency

No relationships were found for consistency on the ba-da continuum and cortical structure metrics. We found several negative relationships with our measure of consistency on the s-sh continuum, however. These included clusters of surface area in the left rostral middle frontal gyrus, the left caudal anterior cingulate (see Table 4.2),

the right superior frontal gyrus, right precuneus, right fusiform, and the right caudal middle frontal gyrus (Table 4.2). We also found a negative relationship between s-sh consistency and volume in a cluster in the right superior frontal gyrus (Table 4.3, Figure 4.1).

Table 4.2: Surface area and s-sh consistency. The r value is the average r values of the vertices in the cluster. Size, coordinates, cluster-wise p values, and anatomical region as defined from the Desikan-Killiany atlas are also indicated.

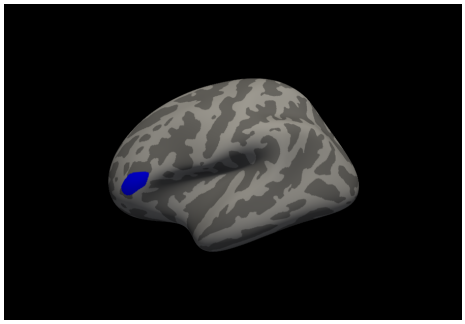
Cluster number	r-value	Size(mm^2)	MNIX	MNIY	MNIZ	p-value	Anatomical region
Left hemisphere							
1	-.47	647.76	-40.4	38.4	0.3	0.0012	rostral middle frontal gyrus
2	-.47	370.44	-7.2	27.6	23.4	0.02623	caudal anterior cingulate
Right hemisphere							
1	-.50	1417.98	14.5	3.2	62.8	0.0002	superior frontal gyrus
2	-.46	708.81	7.7	-61.6	23.5	0.0004	precuneus
3	-.45	376.98	37.1	-38.5	-22.6	0.02445	fusiform gyrus
4	-.45	364.37	29.2	21.7	44.5	0.0284	caudal middle frontal gyrus

Table 4.3: Right hemisphere volume and s-sh consistency. The r value is the average r values of the vertices in the cluster. Size, coordinates, cluster-wise p values, and anatomical region as defined from the Desikan-Killiany atlas are also indicated.

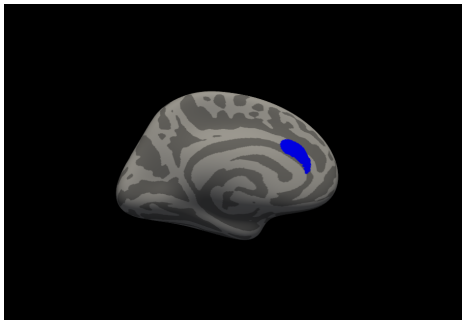
Cluster number	r-value	Size(mm^3)	MNIX	MNIY	MNIZ	p-value	Anatomical region
1	-.47	340.71	19.3	12	61	0.0008	superior frontal gyrus

4.3.4 Region of interest analyses

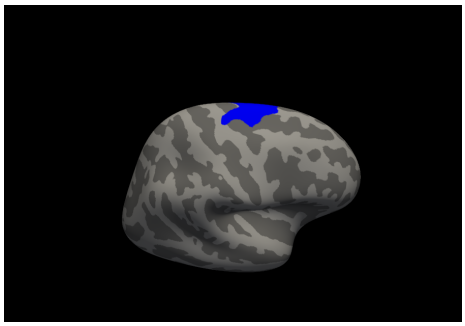
For each of the following analyses, we fit a series of linear regression models that predicted categoricity (slope coefficient) or consistency (mean of squared residuals with the sign changed to facilitate interpretation) for the ba-da and sign-shine continuua. Structural metrics (surface area, cortical thickness, or volume) of each



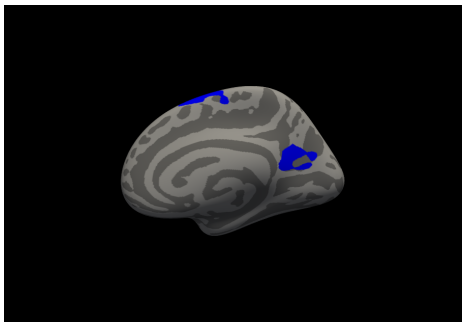
(a) Left hemisphere surface area and s-sh consistency: lateral view.



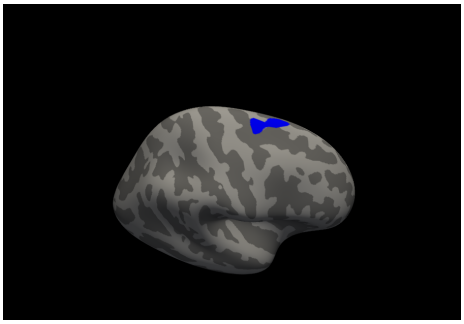
(b) Left hemisphere surface area and s-sh consistency: medial view.



(c) Right hemisphere surface area and s-sh consistency: lateral view.



(d) Right hemisphere surface area and s-sh consistency: medial view.



(e) Right hemisphere volume and s-sh consistency.

Figure 4.1: Clusters predicting response consistency on the s-sh continuum.

region of interest were included as predictors (each in separate models), as well as total intracranial volume when surface area or volume measurements were predictors.

Native-language categoricity

Categoricity ba-da. No structural measurements of the regions of interest predicted the ba-da slope.

Categoricity s-sh. Surface area of the left superior temporal gyrus negatively predicted categoricity when holding other predictors constant, β = -4.682e-05, SE = 2.201e-05, t = -2.13, p = .04, suggesting that more surface area in this region was related to more graded responses on the s-sh continuum. Surface area of the right middle frontal gyrus positively predicted categoricity when holding other predictors constant, β = 2.183e-05, SE = 9.161e-06, t = 2.38, p = .02 (see Figure 4.2), suggesting that individuals with more surface area in this region showed more categorical patterns of perception.

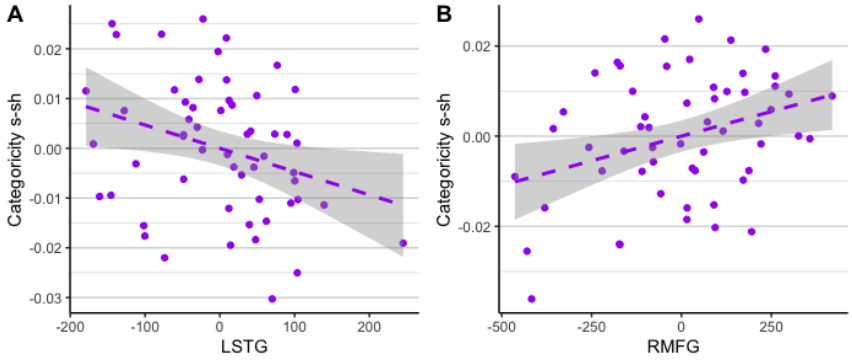


Figure 4.2: Surface area of left superior temporal gyrus (LSTG) and right middle frontal gyrus (RMFG) predicting s-sh slope (categoricity) when holding other predictors constant.

Native-language response consistency

Response consistency ba-da. No structural metrics from our regions of interest predicted response consistency on the ba-da continuum.

Response consistency s-sh. No structural metrics from our regions of interest predicted response consistency on the sign-shine continuum.

4.3.5 Gyrification

Native-language categoricity. To test whether gyrification of the transverse temporal gyri predicted measures of categoricity in the native-language, we fit two linear regression models that predicted categoricity (slope coefficients from visual analog scaling tasks). Fixed effects included only the interaction of local gyrification and hemisphere. This allowed us to test the simple effects of the local gyrificaion index on the dependent variable in each hemisphere separately. Hemisphere was deviation coded as in the previous models. The local gyrification index of either hemisphere did not significantly predict categoricity in either continuum (ba-da or sign-shine).

Response consistency. To test whether gyrification of the transverse temporal gyri predicted measures of response consistency on the native categorization task, we fit two linear regression models that predicted response consistency (the mean of the squared residuals, again with the sign changed to facilitate interpretation). Fixed effects in both models included the interaction of the local gyrification index and hemisphere. The first model predicted response consistency on the ba-da continuum. Local gyrification index in the left hemisphere negatively predicted response consistency, β = -.290, SE = .113, t = -2.573, p = .011, as well as in the right hemisphere, β = -.286,

SE = .112, t = -2.567, p = .012, suggesting that individuals with more gyrification in the transverse temporal gyri are less consistent (or more variable) in their responses on the visual analog scaling task. The second model predicted response consistency on the sign-shine continuum. Local gyrification in the left hemisphere negatively predicted response consistency, β = -.294, SE = .118, t = -2.485, p = .015, as well as in the right hemisphere, β = -.292, SE = .117, t = -2.489, p = .014. This also suggests that participants with more gyrification were less consistent on the categorization task (see Figure 4.3).

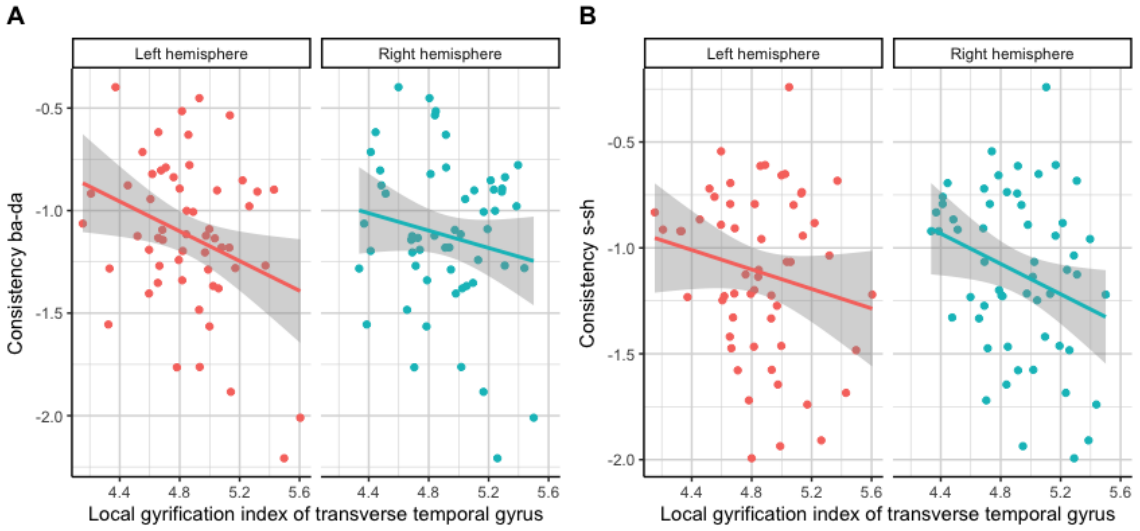


Figure 4.3: Local gyrification index of the bilateral transverse temporal gyri negatively predicts response consistency on the A. ba-da and B. s-sh categorization tasks.

4.4 Discussion

In the current study, we tested whether individual variability in brain structure is related to behavioral performance on native-language speech categorization tasks.

4.4.1 Native-language categoricity

We found very few relationships between brain structure and native-language cateogricity; however, in a region of interest analysis, we found that surface area of the left superior temporal gyrus predicted more gradient responses on a fricative categorization task, and surface area of the right middle frontal gyrus predicted more categorical responses on this task. This parallels previous findings from the functional MRI literature, namely, that the middle frontal gyri (or adjacent regions) show categorical-like responses to native and non-native speech sounds (Luthra et al., 2019; Myers, 2007; Myers et al., 2009; Myers & Mesite, 2014; Myers & Swan, 2012) and that the superior temporal gyrus is sensitive to the graded internal category structure of speech sounds (Myers, 2007). Results from the current study suggest that these findings from the functional literature can be extended to brain structure to predict behavioral performance at the individual level.

4.4.2 Native-language consistency

We tested whether gyrification of the bilateral transverse temporal gyri predicted behavioral measures of native-language response consistency. Several previous studies found that split or duplicate transverse temporal gyri were related to phonetic expertise (Golestani et al., 2011), faster phonetic learning (Golestani et al., 2007), and better non-native speech sound imitation (Turker et al., 2017). Based on these findings, we predicted that more gyrification of the transverse temporal gyrus in either hemisphere would predict more graded perception of native-language speech sounds and more consistent responses on the visual analog scaling tasks. Instead,

we found that gyrification negatively predicted response consistency on the discrete visual analog scaling task, in that participants with more gyrification were less consistent with their responses. This is the opposite of what we predicted based on prior work. Although this apparent discrepancy could have come about from the differences in methodology (i.e., using a continuous measure of gyrification rather than morphological differences in number of gyri), our findings may be relevant to work on the role of categorical perception in reading and language disorders.

Older studies in the field have used two-alternative forced choice tasks to assess categorical perception (e.g., Liberman et al., 1957, see Chapter 1 for a more thorough discussion). Many of these studies have interpreted shallower slopes on an identification task as evidence for more graded perception, and this has been found in younger children (Burnham et al., 1991) and individuals with reading or language disorders (e.g., Serniclaes et al., 2004; Werker & Tees, 1987). However, more recent evidence using eye-tracking suggest that more categorical behavioral response patterns are actually indicative of more graded perception (McMurray et al., 2002) and that perception becomes more graded throughout adolescence (McMurray et al., 2018). This suggests that previous behavioral findings of shallower categorization slopes may be more indicative of noisy representations, which result in less reliable or less consistent responses (McMurray et al., 2002; Kapnoula et al., 2017). In other words, it is possible that the findings from earlier studies showing shallower categorization slopes in individuals with reading and language disorders were actually measuring inconsistent response patterns on speech categorization tasks, rather than truly graded speech category representations. Taking this evidence together with the finding that

more instances of split or duplicate gyri are related to phonological dyslexia (Leonard et al., 2001), gyrification of the transverse temporal gyri seems to predict response consistency in phonetic categorization, even in the typical population. Skoe, Brody, and Theodore (2017) observed variation in the auditory brainstem response that was related to reading ability even among individuals with no history of reading or language disorders, and our results may be reflective of a similar pattern, in which gyrification of the transverse temporaly gyri predicts subtle variation in the typical population (or the broader population more generally) in reading or language ability.

In a whole-brain exploratory analysis, we found several clusters of vertices that negatively predicted response consistency. For example, participants who had increased surface area in the bilateral middle frontal gyri responded less consistently on the fricative categorization (s-sh) task but not the stop (ba-da) categorization task. Although no relationships between brain structure and categoricity emerged in the whole-brain analysis, we would argue that this finding with response consistency is consistent with the idea from the functional MRI literature that the middle frontal gyri and adjacent inferior frontal regions respond categorically to speech stimuli (e.g., Myers et al., 2009; Myers & Swan, 2012; Myers & Mesite, 2014; Luthra et al., 2019). In order to respond consistently on the adapted visual analog scaling task, a listener must be sensitive to the internal structure of speech categories, i.e., they must not only be able to perceive that within category differences exist, but they must also keep track of how close a particular token is to the category boundary throughout the duration of the task. The current findings lend support to the idea that the middle frontal gyri are best suited for detecting changes in speech categories and that

individuals who have increased surface area in these bilateral regions may be less sensitive to the graded internal structure of speech categories.

4.5 Conclusion

The current study explored the structural neural correlates of categorical perception and consistency of responses at an individual level. Many findings reported here complement the functional literature, in that structural measures of frontal regions negatively predicted categoricity and response consistency, while temporal regions showed a positive relationship with categoricity. Local gyrification of the bilateral transverse temporal gyri negatively predicted response consistency, and we speculate that this suggests that gyrification in early auditory regions may be related to subtle variation in language ability in the population.

Chapter 5

General discussion

This dissertation focused on three main issues. First, we attempted to conceptually replicate and extend recent findings involving sleep and overnight improvement on non-native speech sound learning tasks, as well as extend work on structural neural correlates of non-native speech sound learning using surface-based analysis. Second, we tested the hypothesis that how categorically an individual perceives speech sounds in the native language would predict how accurately that individual learns a new non-native speech sound contrast. Finally, we explored the structural neural correlates of categorical perception (i.e., which brain structures would predict how categorically or graded a listener perceived native-language speech sounds).

5.1 Conceptual replication of overnight effects and sleep in non-native speech sound learning

The contributions of sleep and memory consolidation have recently been applied to non-native speech sound learning (e.g., Earle et al., 2017; Earle & Myers, 2015a, 2015b; Fuhrmeister et al., 2020; Qin & Zhang, 2019). Improvement on behavioral tasks even in the absence of further practice is typically taken as evidence of consolidation (e.g. Müller & Pilzecker, 1900) and such improvement is often seen after sleep, but sometimes sleep-dependent consolidation induces qualitative changes to memory representations, even when no changes are seen in behavior (Atienza, Cantero, & Stickgold, 2004; Davis, Di Betta, Macdonald, & Gaskell, 2009; Stickgold, James, & Hobson, 2000, see also Fuhrmeister, 2019; Marshall & Born, 2007, for review). Although there is little doubt that sleep is beneficial for solidifying and retaining newly learned information, several recent studies examining the contributions of sleep-associated memory consolidation to non-native speech sound learning have drawn conclusions based on whether or not *improvement* on a behavioral task was seen after an interval of sleep (e.g., Earle & Myers, 2015a, 2015b; Earle et al., 2017, 2018; Fuhrmeister & Myers, 2017; Fuhrmeister et al., 2020; Qin & Zhang, 2019). Many of these studies were limited in that they had very small sample sizes (13-35 participants per group), and most reported a large amount of individual variability. As discussed in previous chapters, this combination leads to unreliable and unreplicable results. Although there are study design differences, overnight improvement is not found consistently in this literature, and therefore a goal of this dissertation was to

conceptually replicate the finding that learners improve on non-native speech sound learning tasks after an overnight interval and that sleep duration predicts the amount of overnight improvement observed.

With a larger sample size than previous studies ($N = 57$), we assessed non-native speech sound learning in two different tasks, discrimination and identification (the task participants were also trained on) and measured sleep duration during the overnight interval. We found very little evidence that learners improve overnight on the non-native learning tasks, though we observed a non-significant numerical increase in discrimination performance. We also failed to replicate the finding that sleep duration predicts how much a learner will improve overnight. Although some slight differences in study design, stimuli, or analysis approaches could account for the failure to replicate, these results call into question many claims based on overnight improvement (e.g., group differences found in overnight improvement on non-native learning tasks). This does not, however, mean that sleep is not helpful or useful, as there is ample evidence in many literatures that sleep benefits beyond simply improvement on a task (Marshall & Born, 2007). It does, however, bring up the question whether we should be making claims about group differences in overnight improvement if this finding is not reliable or robust. It is problematic for the field that almost all published studies of non-native speech sound learning report a wide range of individual variability and simultaneously have small sample sizes.

If we want to make progress in this area of research, it is essential that we run higher-powered studies to get more precise estimates of effect sizes. One way to do this is to employ statistical approaches that place importance on the precision

of estimates, and data is collected until that precision is reached (e.g., Freedman, Lowe, & Macaskill, 1984, see Kruschke & Liddell, 2018, for a Bayesian approach). Some scientific questions do not lend themselves well to recruiting large sample sizes, such as research with special populations or multi-session studies. In such cases, Bayesian analyses would be more appropriate, as they incorporates prior knowledge into the analysis (e.g., Goodman, 2001). Goodman (2001) argues that even thinking and writing about data with a Bayesian perspective in mind can help mitigate these problems of unreliable results because this way of thinking is more cumulative in nature. In other words, any given data set is only one piece of information that can influence what we already know, and we as researchers should instead ask questions such as how much results from a certain experiment change our beliefs. Even in the context of frequentist statistics, though, the scientific community can take care to interpret results from small sample sizes with caution and to place less value on p-values. This will minimize publication bias so that conclusions drawn from cumulative studies or meta-analyses are less biased.

5.2 Behavioral relationships between native and non-native speech processing

We tested the hypothesis that individuals with more graded native-language speech representations would be more successful at learning non-native speech sounds. This prediction was based on theories of non-native speech sound learning that suggest that assimilating non-native speech sounds to native-language sounds is what makes

perceiving non-native speech sound learning difficult (e.g., Best & Tyler, 2007; Kuhl et al., 2008). We did not find evidence that more graded perception of native speech sounds predicted non-native speech sound learning, which presents a challenge for these theories. It may be, however, that these theories cannot account for difficulty of perception by various individuals, but rather how an individual's native language affects which sounds will be difficult to learn (e.g., voiced dental and retroflex stops are easy to discriminate by native speakers of Hindi but hard for native speakers of English). To be fair, these theories make no claims about perception at the individual level, but rather, they claim that the difficulty perceiving non-native speech sounds stems from their perceptual similarity to native-language speech sounds. However, these theories do not account for the fact that many individuals with the same native language exhibit vastly different patterns of naive perception, learning, and retention of non-native speech sounds. As discussed in Chapter 3, we might need a more sensitive or more automatic measure of sensitivity to within-category differences than we obtained in this project. For example, the degree of phonetic competition (e.g., from varying sounds along a continuum) that an individual experiences as measured by eye tracking or reaction time data could be a more accurate measure of categoricity. It is possible that with a more sensitive measure, we would see relationships between native and non-native speech perception. Currently however, an open question is what cognitive or perceptual processes predict individual variability in non-native speech sound learning among speakers of the same native-language. Many potential contributors to non-native speech sound learning have been explored, such as motivation and language learning aptitude (e.g., Piske, MacKay, & Flege,

2001) or phonological skills (e.g., phonological awareness and phonological working memory, Earle & Arthur, 2017; Fuhrmeister et al., accepted; MacKay et al., 2001; Perrachione et al., 2011), and musical training or ability (Kempe, Bublitz, & Brooks, 2015; Slevc & Miyake, 2006). Though these factors have been shown to explain some amount of individual variability (though not always consistently), we nonetheless lack a robust account of why adult speech sound learners vary so drastically. Access to optimal speech category learning systems seems like a promising avenue for predicting individual differences in speech sound learning (Chandrasekaran, Koslov, & Maddox, 2014; Chandrasekaran, Yi, & Maddox, 2014). Some work has even shown that individuals with elevated depressive symptoms perform better on a speech category learning task than those without depressive symptoms (Maddox et al., 2014). This is due to a deficit in the reflective, non-optimal category learning system, which allows for more reliance on the optimal system because the two systems compete. However, this has only been tested for learning of tonal contrasts, and future work could expand this question to segmental contrasts, as well.

5.3 Structural neural correlates of native and non-native speech

A goal of this dissertation was to extend findings of brain-behavior relationships that have been found for non-native speech sound learning, as well as to test whether structural measurements of certain regions known to be involved with categorical perception in the native language can predict how categorically or graded an individual

perceives speech sounds. Many previous studies examining relationships between brain structure and behavior in the speech perception literature have used voxel-based approaches (e.g., Golestani et al., 2002, 2007) and have reported relationships between behavior and gray matter volume. However, volume measurements are derived from both surface area and cortical thickness. There is recent evidence that cortical thickness and surface area stem from independent genetic processes, and furthermore, that volume is influenced more by surface area than cortical thickness (Winkler et al., 2010). In the current study, we found no relationships between behavior and cortical thickness, only between behavior and volume or surface area. The similar results we found for volume and surface area are consistent with the idea that volume is influenced more by surface area than by cortical thickness, and these findings could be of interest for researchers testing genetic influences on individual or group differences in categorical perception of speech or non-native speech sound learning.

The structural MRI analyses revealed some parallels between native and non-native speech, specifically between the structures that predicted the non-native pretest and native-language response consistency measures. For example, we found in several analyses that surface area in frontal regions (e.g., middle frontal gyrus) negatively predicted performance on the discrimination pretest, as well as response consistency on the categorization task. In addition, we found that gyrification in the transverse temporal gyrus negatively predicted response consistency (more gyrification was related to less consistent responses), and the relationship between gyrification in this region and overnight improvement found for the non-native learning tasks was actually driven by a negative relationship that went away after consolidation. The

fact that we found some parallels between structural relationships with the two tasks suggests that these tasks may share some similarities or may be tapping into a common neural mechanism. From the current data set, it is hard to be more specific about these parallels. Behaviorally, we did not see that response consistency was a strong predictor of non-native speech sound learning. However, the MRI results are suggestive that these tasks may be subserved by some common neural mechanisms.

In general, it is interesting that we found several *negative* relationships between brain structure and behavior because this suggests that "less is more" for certain tasks. Some of these relationships went in the opposite direction of what would be expected from previous literature or from studies of brain function. For example, we predicted based on several previous studies (e.g., Golestani et al., 2007, 2001; Turker et al., 2017) that more gyrification in the transverse temporal gyri would predict better non-native speech sound learning, more graded native-language representations, and more consistent response on the native speech task. As discussed in Chapters 2 and 4, this could reflect a difference in how gyrification was measured in each study, but increased gyrification has also been linked to reading disorders (Leonard et al., 2001; Williams, Juranek, Cirino, & Fletcher, 2018). Thus, more is not always better. In addition, Luthra et al. (2019) found that the middle frontal gyri were sensitive to non-native category differences even before participants were trained on the contrast. Therefore, it seems logical to assume that *more* of a given structural measurement in that area (e.g., surface area, cortical thickness) would result in more accurate pre-training discrimination of a non-native contrast. However, this may be a place where interpreting structural and functional work differs. In studies of

brain function, the goal is usually to understand which structures are involved in performing a certain task, but in studies of brain structure, the goal is typically to understand the relative neural strengths and weakness that contribute to individual or group differences. Therefore, while structure often parallels function, we cannot assume that this is always the case. Perhaps even when a certain brain region is involved in performing a certain task, we may not find that structural variation in this region predicts individual or group performance on the task.

Other findings were more consistent with previous literature. For example, though we found few brain-behavior relationships for individual differences in categorical perception, these seem to parallel findings from functional MRI studies. Specifically, we found that surface area of the right middle frontal gyrus negatively predicted categoricity and that surface area of the left superior temporal gyrus positively predicted categoricity (both relationships were found for the fricative continuum only). This is consistent with previous work showing that the superior temporal gyrus is more sensitive to the graded structure of speech categories (Myers, 2007) and that (inferior) frontal regions are more sensitive to category differences (Luthra et al., 2019; Myers, 2007; Myers et al., 2009; Myers & Mesite, 2014). The current findings suggest that these structures are not only involved in different aspects of speech perception, but variation in their surface area can predict how categorically an individual perceives sounds. We should caution, however, that these relationships were found in a multiple regression model with several other a priori selected regions of interest, which means these relationships were found when holding the other predictors constant. It is therefore entirely possible that, had we chosen slightly

different regions of interest, we would have found different relationships. Nonetheless, surface area of these regions predicted unique variance in categoricity, which adds to our understanding of how these regions contribute to perception of speech sounds.

Another finding that is consistent with previous literature is that hippocampal volume positively predicted overnight change in non-native discrimination performance. This is important because it underscores the importance of memory processes in non-native speech sound learning. Although we did not find significant overnight improvement in behavior alone, it is nonetheless interesting that individuals with more hippocampal volume improved more on the task after sleep. This finding adds support to a growing literature that memory processes are important for non-native speech sound learning, even if behavioral findings that draw conclusions from overnight improvement may need to be reexamined.

Overall, we have some suggestive findings that individual variation in brain structure predicts behavior on native and non-native speech perception tasks. For the non-native speech sound learning tasks, we had tested performance after a delay of about 12 hours (the next day). Except for a weak relationship between volume of the left inferior frontal gyrus and non-native discrimination on the second day, we did not find any relationships with brain structure. In fact, some relationships, for example with transverse temporal gyrus gyrification, even disappeared after a delay. This suggests that brain structure may not be a robust predictor of non-native speech sound learning, which makes it all the more puzzling that we see so much variability in behavior in this process.

One caveat is that these laboratory tasks may not be very reflective of learning

in the real world, and it is an open question whether variation in brain structure can predict naturalistic learning. For instance, Fuhrmeister et al. (accepted) found that adults performed better than children on the same non-native speech sound learning tasks that were tested in this dissertation. This is counterintuitive because we know that children achieve much better real-world language learning outcomes than adults (e.g., Flege et al., 1995, 1999), and it suggests that perhaps the tasks often used in laboratory settings (e.g., identification or discrimination tasks) may not reflect speech sound learning ability in naturalistic settings. One potential way to move forward on this question would be to test whether these laboratory tasks actually predict individual variability in longer-term outcomes of perception and production of non-native languages that are learned in adulthood. The growing use of online data collection platforms could also grant access to more diverse populations of participants and facilitate the collection of larger sample sizes. Much of the current research on language is moving towards testing linguistic phenomena using more ecologically valid stimuli and paradigms, so this question is certainly timely and will be of interest for future work to address.

Appendix: Standardized tests

A.1 Language ability and cognitive skills

Cognitive and language skills have been shown to be related to both native-language speech perception and non-native speech sound learning. For example, many individuals with dyslexia as well as some with specific language impairment have been shown to perceive native-language speech sounds less categorically, as indicated by shallower categorization functions (Joanisse et al., 2000; Serniclaes et al., 2004; Werker & Tees, 1987) or better within-category discrimination (Bogliotti, Serniclaes, Messaoud-Galusi, & Sprenger-Charolles, 2008). Furthermore, Earle, Landi, and Myers (2018) found that individuals with specific language impairment do not show evidence of sleep-mediated consolidation of newly learned non-native speech sounds, whereas those with typical language abilities do. Thus, it is possible that individual differences in reading or language ability might predict outcomes on the native and non-native speech tasks we will ask participants to do. In fact, Earle & Arthur (2017) found that two measures of native-language phonological processing, namely non-word repetition and sound blending, positively predicted learning of a non-native speech sound contrast in adults with and without language impairment. Fuhrmeister et al. (accepted) found that sound blending predicted performance on non-native speech sound learning tasks in a typically developing adult population. Perrachione et al. (2011) additionally found that sound blending predicted learning of a non-native tonal contrast; however, verbal working memory, as measured by a reverse digit span, did not. Even so, verbal working memory or phonological short-term memory is plausibly related to non-native speech sound learning, as holding sounds in memory in order to categorize them may be advantageous to forming long-term memories of these categories (Golestani & Zatorre, 2009; MacKay et al., 2001). MacKay et al. (2001) found that phonological short-term memory as measured by non-word repetition explained a significant amount of the variance in a speech-in-noise task completed by non-native speakers of English. Taken together, these findings suggest

that individual differences in language ability or cognitive skills may be related to native or non-native speech processing.

A.2 Standardized cognitive tests

Because individual differences in language ability and cognitive skills may influence native and non-native speech processing, we administered two tests of native-language phonological processing: the non-word repetition task from the Comprehensive Test of Phonological Processing (CTOPP, Wagner, et al., 1999) and the sound blending task from the Woodcock-Johnson III (WJ-III, Woodcock et al., 2001), and two tests of working memory: auditory working memory, and numbers reversed tasks from the WJ-III (Woodcock et al., 2001). These data are not reported in the main dissertation; however, descriptive statistics are reported below in Table 1.

Data from 53 participants is included here. Five of the 58 participants who were originally recruited were eliminated from these analyses: one for an equipment failure, one for not completing all sessions of the experiment, and three due to experimenter error in administering the standardized tests.

Test	Mean	SD	Minimum	Maximum	Highest possible score
Non-word repetition	9.26	2.54	5	16	18
Sound blending	27.42	3.57	19	33	33
Numbers reversed	16.64	3.69	10	27	30
Auditory working memory	33.23	4.45	22	42	42

Table 1: Descriptive statistics for each standardized test ($N = 53$).

References

Albonico, A., & Barton, J. J. (2017). Face perception in pure alexia: Complementary contributions of the left fusiform gyrus to facial identity and facial speech processing. *Cortex*, *96*, 59–72.

Amrhein, V., Korner-Nievergelt, F., & Roth, T. (2017). The earth is flat (p> 0.05): significance thresholds and the crisis of unreplicable research. *PeerJ*, *5*, e3544.

Atienza, M., Cantero, J. L., & Stickgold, R. (2004). Posttraining sleep enhances automaticity in perceptual discrimination. *Journal of cognitive neuroscience*, *16*(1), 53–64.

Baker, W., Trofimovich, P., Flege, J. E., Mack, M., & Halter, R. (2008). Child—adult differences in second-language phonological learning: The role of cross-language similarity. *Language and Speech*, *51*(4), 317–342.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... Bolker, M. B. (2015). Package 'lme4'. *Convergence*, *12*(1), 2.

Bellander, M., Berggren, R., Mårtensson, J., Brehmer, Y., Wenger, E., Li, T.-Q., ... Lövdén, M. (2016). Behavioral correlates of changes in hippocampal gray matter structure during acquisition of foreign vocabulary. *Neuroimage*, *131*, 205–213.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, *109*(2), 775–794.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by english-speaking adults and infants. *Journal of experimental psychology: human perception and performance*, *14*(3), 345.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. *Language experience in second language speech learning: In honor of James Emil Flege*, *1334*, 1–47.

Bidelman, G. M., Moreno, S., & Alain, C. (2013). Tracing the emergence of categorical

speech perception in the human auditory system. *Neuroimage*, *79*, 201–212.
Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature*, *568*(7753).
Blumstein, S. E., Myers, E. B., & Rissman, J. (2005). The perception of voice onset time: an fmri investigation of phonetic category structure. *Journal of Cognitive Neuroscience*, *17*(9), 1353–1366.
Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer [computer program]. version 5.3. 51. *Online: http://www. praat. org, accessed on*, *2*.
Bogliotti, C., Serniclaes, W., Messaoud-Galusi, S., & Sprenger-Charolles, L. (2008). Discrimination of speech sounds by children with dyslexia: Comparisons with chronological age and reading level controls. *Journal of experimental child psychology*, *101*(2), 137–155.
Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training japanese listeners to identify english/r/and/l: Long-term retention of learning in perception and production. *Perception & psychophysics*, *61*(5), 977–985.
Bradlow, A. R., & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America*, *121*(4), 2339–2349.
Burnham, D. K., Earnshaw, L. J., & Clark, J. E. (1991). Development of categorical identification of native and non-native bilabial stops: infants, children and adults. *Journal of Child Language*, *18*(2), 231–260.
Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.
Chandrasekaran, B., Koslov, S. R., & Maddox, W. T. (2014). Toward a dual-learning systems model of speech category learning. *Frontiers in psychology*, *5*, 825.
Chandrasekaran, B., Yi, H.-G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic bulletin & review*, *21*(2), 488–495.
Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature neuroscience*, *13*(11), 1428.
Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809.
Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *Neuroimage*, *9*(2), 179–194.
Damasio, A. R., & Geschwind, N. (1984). The neural basis of language. *Annual review of neuroscience*, *7*(1), 127–147.

Davis, M. H., Di Betta, A. M., Macdonald, M. J., & Gaskell, M. G. (2009). Learning and consolidation of novel spoken words. *Journal of cognitive neuroscience*, *21*(4), 803–820.

Desai, R., Liebenthal, E., Waldron, E., & Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *Journal of Cognitive Neuroscience*, *20*(7), 1174–1188.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., ... others (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, *31*(3), 968–980.

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage*, *53*(1), 1–15.

Díaz, B., Baus, C., Escera, C., Costa, A., & Sebastián-Gallés, N. (2008). Brain potentials to native phoneme discrimination reveal the origin of individual differences in learning the sounds of a second language. *Proceedings of the National Academy of Sciences*, *105*(42), 16083–16088.

Drouin, J. R., Theodore, R. M., & Myers, E. B. (2016). Lexically guided perceptual tuning of internal phonetic category structure. *The Journal of the Acoustical Society of America*, *140*(4), EL307–EL313.

Earle, F. S., & Arthur, D. T. (2017). Native phonological processing abilities predict post-consolidation nonnative contrast learning in adults. *The Journal of the Acoustical Society of America*, *142*(6), EL525–EL531.

Earle, F. S., Landi, N., & Myers, E. B. (2017). Sleep duration predicts behavioral and neural differences in adult speech sound learning. *Neuroscience letters*, *636*, 77–82.

Earle, F. S., Landi, N., & Myers, E. B. (2018). Adults with specific language impairment fail to consolidate speech sounds during sleep. *Neuroscience letters*, *666*, 58–63.

Earle, F. S., & Myers, E. B. (2014). Building phonetic categories: An argument for the role of sleep. *Frontiers in psychology*, *5*, 1192.

Earle, F. S., & Myers, E. B. (2015a). Overnight consolidation promotes generalization across talkers in the identification of nonnative speech sounds. *The Journal of the Acoustical Society of America*, *137*(1), EL91–EL97.

Earle, F. S., & Myers, E. B. (2015b). Sleep and native language interference affect non-native speech sound learning. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1680.

Eimas, P. D. (1963). The relation between identification and discrimination along

speech and non-speech continua. *Language and Speech*, *6*(4), 206–217.
Finn, A. S., Kharitonova, M., Holtby, N., & Sheridan, M. A. (2019). Prefrontal and hippocampal structure predict statistical learning ability in early childhood. *Journal of cognitive neuroscience*, *31*(1), 126–137.
Fischl, B. (2012). Freesurfer. *Neuroimage*, *62*(2), 774–781.
Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., ... others (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355.
Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, *92*, 233–277.
Flege, J. E., Munro, M. J., & MacKay, I. R. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, *97*(5), 3125–3134.
Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of memory and language*, *41*(1), 78–104.
Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human perception and performance*, *28*(2), 349.
Freedman, L. S., Lowe, D., & Macaskill, P. (1984). Stopping rules for clinical trials incorporating clinical opinion. *Biometrics*, 575–586.
Fuhrmeister, P. (2019). Speech production and perception: Learning and memory. In S. Fuchs, J. Cleland, & A. Rochet-Capellan (Eds.), (p. 207-243). Berlin: Peter Lang.
Fuhrmeister, P., & Fuchs, S. (in preparation). Effects of movement on memory for non-native speech sounds..
Fuhrmeister, P., & Myers, E. B. (2017). Non-native phonetic learning is destabilized by exposure to phonological variability before and after training. *The Journal of the Acoustical Society of America*, *142*(5), EL448–EL454.
Fuhrmeister, P., & Myers, E. B. (2020). Desirable and undesirable difficulties: Influences of variability, training schedule, and aptitude on nonnative phonetic learning. *Attention, Perception, & Psychophysics*, *82*(4), 2049-2065.
Fuhrmeister, P., Schlemmer, B., & Myers, E. B. (accepted). Adults show initial advantages over children learning difficult non-native speech sounds. *Journal of Speech, Language, and Hearing Research*.
Fuhrmeister, P., Smith, G., & Myers, E. B. (2020). Overlearning of non-native speech sounds does not result in superior consolidation after a period of sleep. *The Journal of the Acoustical Society of America*, *147*(3), EL289–EL294.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331.

Golestani, N. (2014). Brain structural correlates of individual differences at low-to high-levels of the language processing hierarchy: A review of new approaches to imaging research. *International Journal of Bilingualism*, *18*(1), 6–34.

Golestani, N., Molko, N., Dehaene, S., LeBihan, D., & Pallier, C. (2007). Brain structure predicts the learning of foreign speech sounds. *Cerebral cortex*, *17*(3), 575–582.

Golestani, N., Paus, T., & Zatorre, R. J. (2002). Anatomical correlates of learning novel speech sounds. *Neuron*, *35*(5), 997–1010.

Golestani, N., Price, C. J., & Scott, S. K. (2011). Born with an ear for dialects? structural plasticity in the expert phonetician brain. *Journal of Neuroscience*, *31*(11), 4213–4220.

Golestani, N., & Zatorre, R. J. (2004). Learning new sounds of speech: reallocation of neural substrates. *Neuroimage*, *21*(2), 494–506.

Golestani, N., & Zatorre, R. J. (2009). Individual differences in the acquisition of second language phonology. *Brain and language*, *109*(2-3), 55–67.

Goodman, S. N. (2001). Of p-values and bayes: a modest proposal. *Epidemiology*, *12*(3), 295–297.

Gow, D. W. (2001). Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory and Language*, *45*(1), 133–159.

Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate l2 attainment in three linguistic domains. *Second Language Research*, *29*(3), 311–343.

Greve, D. N., & Fischl, B. (2018). False positive rates in surface-based anatomical analysis. *NeuroImage*, *171*, 6–14.

Guenther, F. H., Husain, F. T., Cohen, M. A., & Shinn-Cunningham, B. G. (1999). Effects of categorization and discrimination training on auditory perceptual space. *The Journal of the Acoustical Society of America*, *106*(5), 2900–2912.

Hauptmann, B., Reinhart, E., Brandt, S. A., & Karni, A. (2005). The predictive value of the leveling off of within session performance for procedural memory consolidation. *Cognitive Brain Research*, *24*(2), 181–189.

Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of phonetics*, *28*(4), 377–396.

Healy, A. F., & Repp, B. H. (1982). Context independence and phonetic mediation in categorical perception. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(1), 68.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of american english vowels. *The Journal of the Acoustical society of America*, *97*(5), 3099–3111.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLos med*, *2*(8), e124.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, 640–648.

Joanisse, M. F., Manis, F. R., Keating, P., & Seidenberg, M. S. (2000). Language deficits in dyslexic children: Speech perception, phonology, and morphology. *Journal of experimental child psychology*, *77*(1), 30–60.

Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., & McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(9), 1594.

Kempe, V., Bublitz, D., & Brooks, P. J. (2015). Musical ability and non-native speech-sound processing are linked through sensitivity to pitch and spectral information. *British Journal of Psychology*, *106*(2), 349–366.

Köhler, S., Black, S., Sinden, M., Szekely, C., Kidron, D., Parker, J., ... others (1998). Memory impairments associated with hippocampal versus parahippocampal-gyrus atrophy: an mr volumetry study in alzheimer's disease. *Neuropsychologia*, *36*(9), 901–914.

Kong, E. J., & Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, *59*, 40–57.

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic bulletin & review*, *25*(1), 155–177.

Kuhl, P. K. (1994). Learning and representation in speech and language. *Current opinion in neurobiology*, *4*(6), 812–822.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (nlm-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 979–1000.

Lee, Y.-S., Turkeltaub, P., Granger, R., & Raizada, R. D. (2012). Categorical speech processing in broca's area: an fmri study using multivariate pattern-based analysis. *Journal of Neuroscience*, *32*(11), 3942–3948.

Leonard, C. M., Eckert, M. A., Lombardino, L. J., Oakland, T., Kranzler, J., Mohr, C. M., ... Freeman, A. (2001). Anatomical risk factors for phonological dyslexia. *Cerebral cortex*, *11*(2), 148–157.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, *74*(6), 431.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, *54*(5), 358.

Lim, S.-j., & Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive science*, *35*(7), 1390–1405.

Luthra, S., Fuhrmeister, P., Molfese, P. J., Guediche, S., Blumstein, S. E., & Myers, E. B. (2019). Brain-behavior relationships in incidental learning of non-native phonetic categories. *Brain and language*, *198*, 104692.

MacKay, I. R., Meador, D., & Flege, J. E. (2001). The identification of english consonants by native speakers of italian. *Phonetica*, *58*(1-2), 103–125.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.

Maddox, W. T., Chandrasekaran, B., Smayda, K., Yi, H.-G., Koslov, S., & Beevers, C. G. (2014). Elevated depressive symptoms enhance reflexive but not reflective auditory category learning. *cortex*, *58*, 186–198.

Marie, D., Maingault, S., Crivello, F., Mazoyer, B., & Tzourio-Mazoyer, N. (2016). Surface-based morphometry of cortical thickness and surface area associated with heschl's gyri duplications in 430 healthy volunteers. *Frontiers in human neuroscience*, *10*, 69.

Marshall, L., & Born, J. (2007). The contribution of sleep to hippocampus-dependent memory consolidation. *Trends in cognitive sciences*, *11*(10), 442–450.

Mårtensson, J., Eriksson, J., Bodammer, N. C., Lindgren, M., Johansson, M., Nyberg, L., & Lövdén, M. (2012). Growth of language-related brain areas after foreign language learning. *NeuroImage*, *63*(1), 240–244.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). Opensesame: An open-source, graphical experiment builder for the social sciences. *Behavior research methods*, *44*(2), 314–324.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type i error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive psychology*, *18*(1), 1–86.

McMurray, B., Danelz, A., Rigler, H., & Seedorff, M. (2018). Speech categorization develops slowly through adolescence. *Developmental psychology*, *54*(8), 1472.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42.

Miller, J. L. (1997). Internal structure of phonetic categories. *Language and cognitive processes*, *12*(5-6), 865–870.

Müller, G. E., & Pilzecker, A. (1900). *Experimentelle beiträge zur lehre vom gedächtnis* (Vol. 1). JA Barth.

Myers, E. B. (2007). Dissociable effects of phonetic competition and category typicality in a phonetic categorization task: An fmri investigation. *Neuropsychologia*, *45*(7), 1463–1473.

Myers, E. B. (2014). Emergence of category-level sensitivities in non-native speech sound learning. *Frontiers in neuroscience*, *8*, 238.

Myers, E. B., Blumstein, S. E., Walsh, E., & Eliassen, J. (2009). Inferior frontal regions underlie the perception of phonetic category invariance. *Psychological Science*, *20*(7), 895–903.

Myers, E. B., & Mesite, L. M. (2014). Neural systems underlying perceptual adjustment to non-standard speech tokens. *Journal of memory and language*, *76*, 80–93.

Myers, E. B., & Swan, K. (2012). Effects of category learning on neural sensitivity to non-native phonetic categories. *Journal of Cognitive Neuroscience*, *24*(8), 1695–1708.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive psychology*, *47*(2), 204–238.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130*(1), 461–472.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, *24*(2), 175–184.

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2019). R core team. 2019. nlme: linear and nonlinear mixed effects models. r package version 3.1-141. *Available at h Ttp://CRAN. R-Project. Org/Package= Nlme*.

Piske, T., MacKay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an l2: A review. *Journal of phonetics*, *29*(2), 191–215.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across

phonetic categories. *Perception & psychophysics*, *15*(2), 285–290.
Pohlack, S. T., Meyer, P., Cacciaglia, R., Liebscher, C., Ridder, S., & Flor, H. (2014). Bigger is better! hippocampal volume and declarative memory performance in healthy young men. *Brain Structure and Function*, *219*(1), 255–267.
Polka, L. (1991). Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions. *The Journal of the Acoustical Society of America*, *89*(6), 2961–2977.
Price, C. J. (2012). A review and synthesis of the first 20 years of pet and fmri studies of heard speech, spoken language and reading. *Neuroimage*, *62*(2), 816–847.
Qin, Z., & Zhang, C. (2019). The effect of overnight consolidation in the perceptual learning of non-native tonal contrasts. *PloS one*, *14*(12).
Qu, J., Zhang, L., Chen, C., Xie, P., Li, H., Liu, X., & Mei, L. (2019). Cross-language pattern similarity in the bilateral fusiform cortex is associated with reading proficiency in second language. *Neuroscience*, *410*, 254–263.
Rakic, P. (2000). Radial unit hypothesis of neocortical expansion. In *Novartis foundation symposium* (pp. 30–52).
Repp, B. H. (1981). Two strategies in fricative discrimination. *Perception & Psychophysics*, *30*(3), 217–227.
Rodriguez, S. M., Archila-Suerte, P., Vaughn, K. A., Chiarello, C., & Hernandez, A. E. (2018). Anterior insular thickness predicts speech sound learning ability in bilinguals. *Neuroimage*, *165*, 278–284.
Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038.
Schaer, M., Cuadra, M. B., Schmansky, N., Fischl, B., Thiran, J.-P., & Eliez, S. (2012). How to measure cortical folding from mr images: a step-by-step tutorial to compute local gyrification index. *JoVE (Journal of Visualized Experiments)*(59), e3417.
Schremm, A., Novén, M., Horne, M., Söderström, P., van Westen, D., & Roll, M. (2018). Cortical thickness of planum temporale and pars opercularis in native language tone processing. *Brain and language*, *176*, 42–47.
Sebastián-Gallés, N., Soriano-Mas, C., Baus, C., Díaz, B., Ressel, V., Pallier, C., ... Pujol, J. (2012). Neuroanatomical markers of individual differences in native and non-native vowel perception. *Journal of Neurolinguistics*, *25*(3), 150–162.
Serniclaes, W., Van Heghe, S., Mousty, P., Carré, R., & Sprenger-Charolles, L. (2004). Allophonic mode of speech perception in dyslexia. *Journal of experimental child psychology*, *87*(4), 336–361.
Shibata, K., Sasaki, Y., Bang, J. W., Walsh, E. G., Machizawa, M. G., Tamaki,

M., ... Watanabe, T. (2017). Overlearning hyperstabilizes a skill by rapidly making neurochemical processing inhibitory-dominant. *Nature neuroscience*, *20*(3), 470–475.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359–1366.

Singmann, H., Bolker, B., & Westfall, J. (2019). *Aust, f. afex: Analysis of factorial experiments. r package version 0.23–0.*

Skoe, E., Brody, L., & Theodore, R. M. (2017). Reading ability reflects individual differences in auditory brainstem function, even into adulthood. *Brain and language*, *164*, 25–31.

Skoe, E., & Kraus, N. (2010). Auditory brainstem response to complex sounds: a tutorial. *Ear and hearing*, *31*(3), 302.

Slevc, L. R., & Miyake, A. (2006). Individual differences in second-language proficiency: Does musical ability matter? *Psychological science*, *17*(8), 675–681.

Stickgold, R., James, L., & Hobson, J. A. (2000). Visual discrimination learning requires sleep after training. *Nature neuroscience*, *3*(12), 1237–1238.

Stölten, K., Abrahamsson, N., & Hyltenstam, K. (2015). Effects of age and speaking rate on voice onset time: The production of voiceless stops by near-native l2 speakers. *Studies in Second Language Acquisition*, *37*(1), 71–100.

Turker, S., Reiterer, S. M., Seither-Preisler, A., & Schneider, P. (2017). "when music speaks": Auditory cortex morphology as a neuroanatomical marker of language aptitude and musicality. *Frontiers in psychology*, *8*, 2096.

Van Breukelen, G. J. (2006). Ancova versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of clinical epidemiology*, *59*(9), 920–925.

Van Petten, C. (2004). Relationship between hippocampal volume and memory ability in healthy individuals across the lifespan: review and meta-analysis. *Neuropsychologia*, *42*(10), 1394–1413.

Venables, W., & Ripley, B. (2002). *Mass library of functions.*

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (1999). *Comprehensive test of phonological processing: Ctopp.* Pro-ed Austin, TX.

Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: new directions. *Annual review of psychology*, *66*, 173–196.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, *7*(1), 49–63.

Werker, J. F., & Tees, R. C. (1987). Speech perception in severely disabled and

average reading children. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *41*(1), 48.

White, T., Su, S., Schmidt, M., Kao, C.-Y., & Sapiro, G. (2010). The development of gyrification in childhood and adolescence. *Brain and cognition*, *72*(1), 36–45.

Wierenga, L. M., Langen, M., Oranje, B., & Durston, S. (2014). Unique developmental trajectories of cortical thickness and surface area. *Neuroimage*, *87*, 120–126.

Williams, V. J., Juranek, J., Cirino, P., & Fletcher, J. M. (2018). Cortical thickness and local gyrification in children with developmental dyslexia. *Cerebral Cortex*, *28*(3), 963–973.

Winkler, A. M., Kochunov, P., Blangero, J., Almasy, L., Zilles, K., Fox, P. T., ... Glahn, D. C. (2010). Cortical thickness or grey matter volume? the importance of selecting the phenotype for imaging genetics studies. *Neuroimage*, *53*(3), 1135–1146.

Wong, P. C., Warrier, C. M., Penhune, V. B., Roy, A. K., Sadehh, A., Parrish, T. B., & Zatorre, R. J. (2008). Volume of left heschl's gyrus and linguistic pitch learning. *Cerebral cortex*, *18*(4), 828–836.

Woodcock, R. W., McGrew, K. S., Mather, N., et al. (2001). Woodcock-johnson iii tests of achievement [Computer software manual]. Riverside Publishing Company Itasca, IL.

Yi, H.-G., Maddox, W. T., Mumford, J. A., & Chandrasekaran, B. (2016). The role of corticostriatal systems in speech category learning. *Cerebral Cortex*, *26*(4), 1409–1420.

Zatorre, R. J., Fields, R. D., & Johansen-Berg, H. (2012). Plasticity in gray and white: neuroimaging changes in brain structure during learning. *Nature neuroscience*, *15*(4), 528–536.